SPEECH KNOWLEDGE MODELLING FOR SPEECH RECOGNITION:

A Study Based on Distinctive Features

Shuping RAN

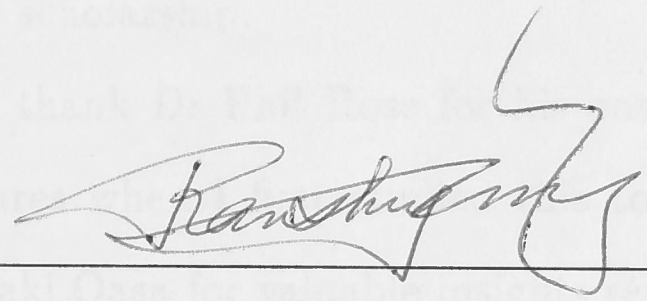
A thesis submitted for the degree of
Doctor of Philosophy
of the Australian National University

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Canberra, Australia

January 1994

Statement

I hereby testify that, except where acknowledgements are given, this thesis is entirely my own original work.



14/1/1994

(Shuping RAN) (date)

Acknowledgements

I would like to thank Drs Bruce Millar and Iain Macleod for their valuable guidance and encouragement, and for their constructive suggestions regarding improvements to draft versions of this thesis.

I acknowledge with gratitude the support of an Australian National University PhD scholarship.

I thank Dr Phil Rose for his enthusiastic discussions concerning distinctive features when I first studied this topic. I also thank my former colleague Dr Hiroaki Oasa for valuable insights regarding phonetics.

I would also like to thank Professor Kenneth Stevens for the useful discussions at the Fourth Australian International Conference on Speech Science and Technology in December 1992.

Thanks are also due to Mr Andrew Tridgell for discussions in computing related matters and to Mr Bill Alford for help with Latex. I thank Dr Frantz Clermont for encouragement and for help in generating the required thesis format. I thank Mr Dennis Andriolo, Mr Joe Elso and Mr Arthur McGuffin for their kindness and constant technical support. I would also like to thank all members of the Computer Sciences Laboratory for their friendship during the last four years.

My special thanks go to my husband Weijia JIA, for his love, understanding, encouragement and support.

Abstract

There are two main approaches to automatic speech recognition: data-driven pattern recognition and knowledge-based techniques. As yet, neither of these approaches has resulted in generally acceptable performance. This lack of success is in part a consequence of our incomplete understanding of speech production and perception, meaning that we are unable to explicate the speech code satisfactorily. This thesis adopts an approach in which data-driven pattern recognition and knowledge-based approaches are integrated in a parallel fashion. Relevant systematic knowledge is encoded in the overall system structure; pattern recognition techniques are used to cope with those aspects where appropriate knowledge is either not available or not expressible in computable form.

The principal form of systematic knowledge incorporated is Jakobson, Fant and Halle's distinctive features. The data-driven pattern recognition technique used is artificial neural networks of the Multi-Layer Perceptron type, which are also used to perform phonetic recognition on the basis of distinctive feature detection results.

By analysing the distinctive feature detection results, we have found that the detectors provide a consistent mapping from acoustic space to distinctive feature space. The inter-vowel relationships in articulatory space are preserved in this distinctive feature space.

We conclude that our approach of integrating the relevant speech knowledge at the system structure level and using pattern recognition techniques to cope with detailed acoustic analysis (in the absence of computable acoustic specifications) can enhance system performance and provide an environment in which further

understanding of the speech code can be gained. Our results show that under this paradigm, a consistent mapping from an acoustic to an articulatory domain can be performed, with potential for objective assessment of the articulatory quality of vowels.

Contents

Chapter 1

Introduction

1.1 Motivation

Because speech is the most natural and efficient means of human communication, human-machine interaction using speech has been a long sought goal. Speech output by machine is relatively satisfactory: the main problem which remains to be solved here is *automatic speech recognition* by machine (ASR). After many years of active research, two major approaches have emerged: knowledge-based approaches and data-driven pattern recognition approaches. However, neither of these approaches has achieved performance comparable to human listeners and it is still unclear which has the best potential for achieving such performance. The difficulty of the problem is due to our incomplete understanding of the processes through which phonetic information is encoded in the speech signal and then decoded by the listener back into phonetic information. An automatic speech recognition system receives as input only the acoustic signal, which encodes several different kinds of information, such as phonetic information, speaker characteristics, and environmental properties. There is little guidance as to how humans decode this complex signal. To date, it has not been possible to automatically separate these different types of information from the acoustic signal. It is however, reasonable to hypothesise that there are sub-structures in the speech signal which are mainly dependent on phonetic information, and that these can in principle be separated out.

Alongside the development of this hypothesis, some initial studies were performed on very simple acoustic structures (Ran and Millar, 1990, 1991a, 1991b, 1991c) which used artificial neural networks to:

- refine rough hand-segmentation of the vocalic and nonvocalic portions of a speech signal.
- perform vocalic and nonvocalic classifications.
- distinguish between transitional and steady-state portions of vocalic type signals.
- recognise vowels and stop consonants using steady-state vocalic signals and restricted portions of the nonvocalic signals respectively.

The pilot study results showed that artificial neural networks (ANNs) could reliably distinguish between vocalic and nonvocalic, transitional and steady-state speech signals. Recognition results using such pre-classified portions of the speech signal were better than those obtained with the complete signal.

The results of these pilot studies tended to confirm the hypothesised existence of phonetically relevant structures (as referred to in the phonetic, linguistic, and perceptual literature) in the acoustic signal. If methods can be found to extract these structures, they should be very helpful for speech recognition.

Traditionally, knowledge-based speech recognition and data-driven pattern recognition techniques are treated separately or at best sequentially. In recent years, there has been a trend to abandon knowledge-based approaches in favour of pattern recognition techniques. This trend is a consequence of our incomplete understanding of the "speech code" and the difficulties in quantifying what knowledge we have in a way that is applicable with existing knowledge-based techniques. Should we continue with the pure pattern recognition technique or can we profit by incorporating certain forms of knowledge in the ASR system? Is it possible to integrate knowledge-based and pattern recognition techniques

in ASR in a parallel fashion for robust speech recognition where the knowledge-based component captures those aspects of the speech code that we can explicate, and data-driven pattern recognition is used for the remainder?

In our pilot studies, we were concerned with only a few very basic structures in the acoustic signal; there is a need to look beyond these simple structures to structures which are more directly relevant phonetically and which at the same time can be reliably extracted. To the extent that we are successful in this quest, the results should be very useful for speech recognition.

We expect that under this integrated approach to speech recognition, relevant systematic knowledge can be integrated in the system structure and the underlying acoustic information need not be specified explicitly as in traditional knowledge-based approaches, but can be extracted automatically via pattern recognition techniques. This paradigm should establish a framework within which the characteristics of system performance can be more readily understood and provide an environment in which details of the speech code can be explored.

1.2 Thesis Outline

This thesis comprises six core chapters followed by a concluding discussion and an appendix.

Chapter 2 constructs a conceptual framework for the research. It reviews automatic speech recognition chronologically, analyses different approaches taken over the past decades of research, and concludes that a combination of knowledge-based and data-driven pattern recognition techniques is a promising approach for automatic speech recognition.

Chapter 3 examines the background theory, starting with a review of one systematic body of speech knowledge — *distinctive features*. It then analyses the invariance and variability issue which is directly related to questions regarding the existence of acoustic correlates of the distinctive features. It finally reviews research on modelling of sub-phonetic units.

Chapter 4 details the selection of the techniques and the phonetic features incorporated in this thesis study, before reviewing Jakobson *et al.*'s distinctive feature theory and artificial neural networks in more detail. A method of feature definition for a relevant subset of Australian English is also described.

Chapter 5 describes the rationale used in selecting the speech material for this study, the criteria for acoustic signal segmentation and labelling, and the acoustic analysis procedure.

Chapter 6 reports the results of automatic speech recognition based on distinctive feature detection, in which the outputs of feature detection modules are used as inputs to a speech recognition module. These results are compared with those from two monolithic approaches where no prior detection of distinctive features is performed.

Chapter 7 analyses the feature detectors' responses for vowels in order to verify some properties of the distinctive features, and to investigate relationships between the feature space generated by the feature detectors' responses and articulatory spaces.

Chapter 8 draws overall conclusions from the work and discusses broader issues.

The Appendix presents detailed supplementary material on which the analysis of Chapter 7 is based.

Chapter 2

Automatic Speech Recognition: A Review

2.1 Introduction

Automatic speech recognition (ASR) is a developing area which has attracted an extensive research effort contributing to several different ASR technologies. This chapter first presents a chronological view of ASR research. It then reviews four schools of thought regarding ASR technologies, followed by a discussion of current problems in ASR and the importance of a knowledge–based approach. The chapter concludes by arguing that phonetic features can usefully be incorporated at the acoustic–phonetic level of ASR.

2.2 A Chronological View

As Ainsworth (1988) observes, the history of automatic speech recognition (ASR) is mainly a catalogue of the realisation of problems involved in ASR. Following Ainsworth, we can summarise the development in the ASR field under five ordered epochs.

2.2.1 Pre–1960: The Acoustic Approach

Because of successful experience using sound spectrographs, the hypothesis that all the information required for recognising speech resided in patterns of spectral energy in the acoustic signal became commonly accepted. First attempts to build recognisers based on acoustic patterns used filter banks for the acoustic analysis.

A reference pattern was formed for each word and the test pattern was cross–correlated with each of the stored reference patterns (Davis *et al.*, 1952; Dudley and Balashek, 1958). The problem with these systems is that they work well with the speaker who generated the patterns, but are much less successful with other speakers.

2.2.2 1960–1968: The Pattern Recognition Approach

By 1960 it was generally realised that the technique of simple matching of acoustic patterns had limited prospect of success. The acoustic patterns of a word repeated by the same speaker change with time, and in duration and intensity; when repeated by different speakers, the spectral content may vary. The idea of normalisation pre-classification was introduced, such as formant frequency normalisation using fundamental frequency (Forgie and Forgie, 1959) and duration normalisation (Denes and Mathews, 1960).

2.2.3 1969–1976: The Linguistic Approach

A basic requirement for a listener to understand a speaker is that they speak the same language. Early attempts at speech recognition mostly neglected such linguistic knowledge, but subsequently the need to incorporate this knowledge was gradually realised.

The many sources of linguistic knowledge may be sub-divided into seven areas:

Acoustic–phonetic knowledge – This represents the relationship between the acoustic signal and a phonetic transcription of the corresponding spoken utterance. This has been incorporated in almost all systems.

Lexical Knowledge – This defines the composition of each word in the vocabulary. Early incorporation of this kind of knowledge was reported by Green and Ainsworth (1972).

Phonological Rules – These represent the systematic changes of pronunciation of words or phonemes according to the environment, as incorporated in an ASR system by Oshika *et al.* (1974) and Shoup (1980).

Prosodic Knowledge – This can be used to locate the stressed syllables in an utterance and to segment the utterance into syntactic phrases. Lea (1973) investigated prosodic features and how could they be used in ASR.

Syntactic Knowledge – This defines how the phrases and sentences are formed. It was first incorporated in ASR by Tappert (1974).

Semantic Knowledge – This helps to form meaningful sentences (Woods, 1974).

Pragmatic Knowledge – This helps to define the likely set of utterances, such as in the system *Hearsay* (Reddy *et al.*, 1973).

Problems were found when integrating different knowledge sources into one system. The simplest integration is bottom up in a hierarchical manner. The acoustic–phonetic processor produces a phonetic lattice which is passed to the lexical processor. This produces a network of possible word matches which is subsequently pruned by the syntactic processor, semantic processor, and the pragmatic processor respectively, in order to select an utterance which is grammatically correct and meaningful in the current context. This integrating method was found to be inefficient, because many of the hypotheses generated at the bottom level were subsequently pruned by later processors. Other methods were proposed and tested, such as the *blackboard* (Reddy *et al.*, 1973) and the network approach (Baker, 1975b; Lowerre, 1976). In the *blackboard* approach, each processor can access (read/write) suggestions to/from a common structure – the *blackboard* – in order to process and refine their own decisions. This approach was used in the Hearsay I and II systems (Reddy *et al.*, 1973; Lesser *et al.*, 1975). Conversely, in the network approach the knowledge sources are integrated via a network. Recognition consists of searching for the path through this network which is most consistent with the input acoustic data. This approach was used in the *Dragon* system (Baker, 1975b) and in the *Harpy* system (Lowerre, 1976).

2.2.4 1977–1986: The Pragmatic Approach

During the decade 1977–1986, work on ASR increased significantly. On one hand, use of Dynamic Programming (DP) algorithms for isolated word recognisers become a general approach after Velichko and Zagoruyko (1970) first used DP. Later, the power of DP to overcome non-linear scaling differences in time was demonstrated by Sakoe and Chiba (1978) and it has been a popular technique since then. On the other hand, to deal with incomplete or uncertain information, an algorithm based on stochastic models which is more general and has a firmer mathematical foundation was also developed during this time and had encouraging results (Jelinek, 1976). Aspects based on hidden Markov modelling were developed by Levinson *et al.* (1983).

During this period, work based on linguistic knowledge continued, such as studies by Klatt (1979), Mercier *et al.* (1980), and Nieman (1982).

2.2.5 1987–1990s: Recent developments

During the last few years, research based on stochastic models (especially that based on hidden Markov models) has become very popular because of the ability of such models to deal with incomplete or uncertain information of speech. Further development of ASR based on this technique has produced improved recognition results, as reported by Lee *et al.* (1989), Lee (1990) and Huang and Jack (1989).

At the same time, interest in Artificial Neural Networks (ANNs), partially motivated by neuroscience, has renewed. Application of different types of ANNs to ASR has been studied, such as Waibel *et al.*'s Time Delay Neural Network (1989) which has a multilayer architecture with time delayed connections in order to handle time shifts and the temporal nature of speech, Kohonen's Self-Organising Map (1988) which conducts an unsupervised search for problem solutions and Sakoe *et al.*'s Dynamic Programming Neural Network (1989) which combines neural networks and Dynamic Programming techniques in order to deal with spectral pattern variations and time axis distortions of speech respectively.

2.3 Different Approaches for ASR

2.3.1 Introduction

Section 2.2 summarised development of automatic speech recognition chronologically. This development has been largely a process of realisation of the difficulty of ASR and of the different levels of complexity in speech signals and consequent processing adjustments made to cope with the growing awareness of the nature and extent of such problems. Without attempting to be comprehensive, the current section reviews techniques used in ASR systems placing emphasis on basic techniques employed: more comprehensive reviews can be found in Mariani (1989), Waibel and Lee (1990), and Furui and Sondhi (1991).

Current techniques for ASR can be divided into two main approaches: knowle dge–based speech recognition, and pattern recognition techniques. Knowledge–based techniques employ a set of rules based on a set of features. These rules are typically derived from human experts. No reference patterns are stored: recognition is performed by applying the rules to the test patterns. In the pattern recognition approach, reference patterns are stored internally as patterns or models representing the reference patterns. Recognition is performed by comparing the test patterns with these stored patterns. Pattern recognition can be divided further into three approaches: Template Matching, Stochastic and Connectionist. The following subsections discuss each of these approaches further.

2.3.2 Template–Based Approach

The template–based approach was the first to be tried (Davis *et al.*, 1952): encouraging results were obtained, initial progress was good and the technique improved very rapidly. This approach has matured after two decades of research effort and a family of techniques has been developed which has contributed greatly to the overall advance of the ASR field.

The basic idea is simple. A set of prototype patterns is created and stored as reference patterns (the "templates") which represent the allowed words in the

word vocabulary. Recognition is carried out by matching an unknown pattern with each of the templates in the word vocabulary. The unknown pattern is assigned to the class of the reference pattern which gives the best match. This technique is mostly applied to word recognition, where each template and unknown isolated token represents a word.

Word-template matching does not need the phonetic segmentation and labelling required in approaches based on smaller units such as phonemes. Although segmentation and labelling are needed at the word level, this needs far less work than phonetic segmentation and labelling do, saving time and avoiding errors resulting from incorrect segmentation and labelling. One shortcoming is that a large memory is required as a full template has to be created for each word in the vocabulary. As the vocabulary size increases, so does the memory demand. Rabiner and Levinson (1981) give a good review of this technique.

During the process of development of this approach, more problems were realised, and the need for methods to overcome them was clarified.

One of the problems was the need to accommodate different speaking rates. A good match between the test pattern and the reference pattern was difficult to obtain if the time course of each was different. Dynamically stretching or compressing the patterns along the time axis was proposed by Sakoe and Chiba (1978) using a Dynamic Programming technique (Bellman, 1957) to achieve a time alignment (which overcomes overall variations over time and the internal variation caused by different speaking rate, speaking style, etc.) Figure 2.1 illustrates the Dynamic Time Warping (DTW) concept where a test pattern ($T(n)$) is compared with a reference pattern ($R(m)$). The best path of match is found by minimising a cost defined as the cumulative distance between each sequential pair of vectors (one of ($T(n)$) and the other of ($R(m)$)) along the path of match in an $n * m$ space.

Another problem realised was the high level of inter-speaker variance which made speaker–independent recognition difficult. Rabiner *et al.* (1979) proposed a

clustering technique to automatically generate multiple templates for each word as a solution to this problem.

A further problem concerned the limitation to isolated word or phrase recognition. Extensions to handle template matching for connected word recognition were proposed by several authors. Sakoe (1979) proposed a two-level DP matching algorithm which used a second-level dynamic–programming optimisation to find the best sequence of words given the optimal word time alignment for each possible word start and end time. Myers and Rabiner (1981) proposed a level-building algorithm which iteratively extends N-word hypotheses to $N+1$ words. Vintsyuk (1971), Bridle and Brown (1979), Bridle *et al.* (1982) and Ney (1984) proposed a one-stage strategy which does not require several levels of optimisation. It relies basically on parameterising the time warping path by a single index and on exploiting certain path constraints both in the word interior and at the word boundaries.

In summary, the template–based approach is a simple technique which is used mostly for speech recognition having words as speech units. This technique has certain limitations discussed above which limit its domain of application: it does not provide a suitable basis for further advances.

2.3.3 Stochastic Approach

The most popular stochastic approach is hidden Markov modelling (HMM). Baker (1975a,b) and Jelinek (1976) first introduced the use of HMM for speech recognition. A good introduction to HMMs is given by Rabiner (1989), and Rabiner and Juang (1986). Recently, Huang *et al.* (1990) have described HMMs in greater detail.

Hidden Markov models use a Markov process (Markov, 1913) to model changing statistical characteristics that are only probabilistically manifested in actual observations. Such processes are characterised by steady periods represented as *states* and transitions from one state to the next represented as *transitions*. Figure 2.2 gives an example of a HMM in which each state has a self loop plus a

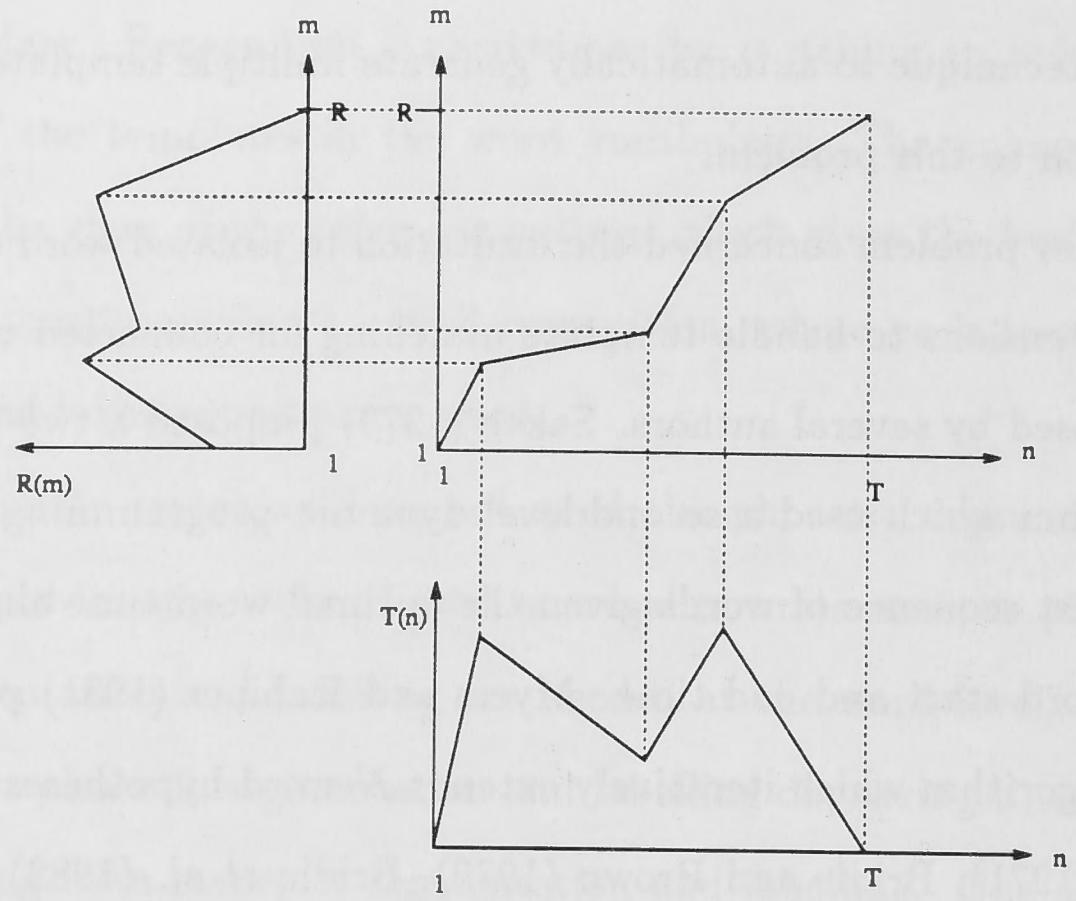


Figure 2.1: An example of dynamic time warping.

transition (*link*) to each of the following states. The term *hidden* refers to the fact that the state of the process modelled by an HMM at time t is not directly observable. The parameters to be determined for an HMM are *transition probabilities* and *output probabilities*. *Transition probabilities* a_{ij} are associated with the link going from state i to state j, representing the probability that this path could be taken. *Output probability* represents the probability of producing an output S_o when the path from state i to state j is taken.

An HMM uses a probabilistic approach to deal with uncertain and/or incomplete information. In speech recognition, uncertainty and incompleteness come from many sources such as speaker variability, confusible sounds, contextual effects, recording conditions, background noise, etc. Thus HMM approaches have been very attractive for ASR applications. Generally speaking, in hidden Markov modelling a speech unit (i.e. word, phone, diphone, triphone, etc.) is modelled by a sequence of states with connections between them. The transition parameters of the HMM attempt to capture the temporal variability, and the output probability parameters model the spectral variability. The number of states, the number of links, and the initial and final states for each link are chosen by the system

designer according to the nature of the problem. The parameters of the model (transition and output probabilities) have to be obtained through "training".

Three problems have to be addressed (Rabiner, 1989; Mariani, 1989):

- *Training* – answers the question of how to get the parameters of the model, given a sequence of labels. This can be obtained by the *forward–backward* (also called Baum–Welch) algorithm (Baum, 1972), when the training is based on maximum likelihood.

- *Decoding* – answers the question of which sequence of states has produced the sequence of labels. This can be obtained by the Viterbi algorithm (Viterbi, 1967), which is very similar to Dynamic Time Warping (DTW).

- *Evaluation* – answers the question of what is the probability that a sequence of labels has been produced by a given model. This can be obtained by using the *forward* algorithm, which gives the maximum likelihood estimation that the sequence was produced by the model.

In this stochastic approach to ASR, the reference patterns are represented by hidden Markov models, whereas in a template–based approach, they are represented by templates stored in the memory. The HMM approach provides ASR with a more flexible representation of the reference patterns and the ability to cope with speaker variabilities in duration and spectral information.

As pointed out in Section 2.2, the history of speech recognition has been a process of realisation of the problems involved, and a subsequent search for solutions to these problems. Such a process can also be observed in the development of HMMs.

Early attempts to apply HMMs to speech recognition used *discrete* HMMs, in which real–valued acoustic parameters were first quantified into discrete symbols. This reduction process sacrifices accuracy for simplicity. To overcome this problem, Poritz and Richter (1986) and Paul *et al.* (1986) proposed modelling

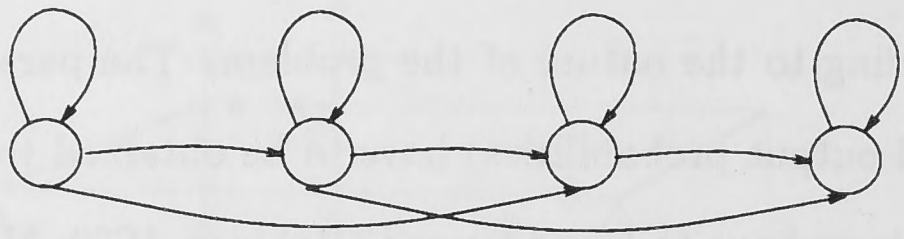


Figure 2.2: An example of a four state left–right HMM.

the output distribution as a mixture Gaussian density, starting the era of *continuous* HMMs. Rabiner *et al.* (1989) reported their continuous HMM system for recognising connected digits. Their results showed much better performance than discrete HMMs. More examples are given in Mariani's (1989) review paper.

Use of continuous HMMs involves considerable computational complexity and is very sensitive to initial estimates of several model parameters (Juang and Rabiner, 1985; Rabiner *et al.*, 1985). To overcome this problem, Huang and Jack (1989) proposed the use of *Semi–Continuous* HMMs. This is a combination of continuous and discrete HMMs, replacing discrete output probabilities with a combination of the original discrete output probability and continuous density functions of a mixture Gaussian code book (Huang and Jack, 1989). Semi–continuous HMMs have been shown to perform better than discrete and continuous HMMs by Huang and Jack (1989), and Bellegarda and Nahamoo (1989).

In summary, although use of HMM technology has contributed to advances in automatic speech recognition, as Rabiner (1989) indicated, there are certain drawbacks to the use of HMMs for speech recognition due to intrinsic limitations of HMMs. One of these is that according to the Markov assumptions, each observation (a frame of speech) is assumed to be independent, and dependent only on the associated state. Another limitation is the Markov assumption namely, that the probability of being in a given state at time t depends only on the state at time $t-1$. These two assumptions are not totally appropriate for speech sounds where adjacent frames are closely related and the probability of being in a given state at a given time often depends on several proceeding states (Rabiner, 1989). Finally, HMMs represent the speech signal by using mathematical models internally, but the mapping between the speech signals and the mathematical

models is not clear. Therefore, HMMs provide little insight into the recognition process. As a result, it is often difficult to analyse the errors of an HMM system when attempting to improve its performance.

2.3.4 Connectionist Approach

In the connectionist approach, a model is represented by a set of nodes (or processing units) and links between them. Each node is associated with a function whose parameters are related to the input to the node. Each link is associated with a weight representing the strength of the link.

As a result of the analogy with the anatomy of neural structures, the *processing units* are sometimes regarded as "neurones" and the *links* as "synapses", and complete architectures are referred to as *Artificial Neural Networks (ANNs)*, or Neural Network for short. From a different analogical starting point, they are also known as *parallel distributed processing systems* due to the fact that they have many simple distributed processing units working in parallel.

As in template–based and stochastic approaches in the connectionist approach, knowledge or constraints are not encoded in the system explicitly, but are distributed across many simple computing units, which are connected to form a structure. Uncertainty is not modelled explicitly by likelihood or probability as in HMMs, but by patterns of activity in many processing units and their interconnecting links. The parameters to be estimated for a connectionist model are the weights representing the strength of the links between the processing units.

The ancestor of this approach is the *perceptron* which was intended as a model of visual perception (Rosenblatt, 1958, 1959). The perceptron was finally abandoned after having been theoretically proved to be incapable of making certain simple discriminations (Minsky and Papert, 1969). Multi-Layer Perceptrons (MLPs) are feed-forward networks with one or more layers of nodes between the input and output nodes. These additional layers contain hidden units or nodes that are not directly connected to both the input and output nodes. MLPs were

generally not used in the past because effective training algorithms were not available. After nearly twenty years, Rumelhart *et al.* (1986b) and others (Werbos, 1974; Le Cun, 1985; Parker 1985) proposed a new and effective training algorithm for MLPs called *Back-Propagation* which made the use of MLPs possible. Interest in using MLPs has been renewed since Lippmann (1987) proved that Multi–Layer Perceptrons (MLP) have superior classification ability. Good introductions and reviews of ANNs can be found in Lippmann (1987) and Rumelhart *et al.* (1986b). Hush and Horne (1993) review more recent developments in the ANN area.

Applications of ANNs to speech recognition are becoming more and more widespread. Connectionist modelling of speech is however still in its infancy and the subject of much controversy (see Section 4.5.6 and below). The development of speech recognition research using ANNs has been reviewed by Lippmann (1989).

In the connectionist approach, reference patterns of speech units are stored in the network in a distributed way. The model parameters (weights on each internode link) are estimated by making use of a training algorithm while repeatedly presenting the reference patterns to the input of the network.

A more detailed description of ANNs is given in Section 4.5, but some general observations can already be made.

Several reasons why ANNs have been attractive for ASR are:

- ANNs learn from the patterns presented to them.
- ANNs are trained to best classify patterns belonging to each class and best discriminate patterns belonging to different classes.
- ANNs learn from examples that include incomplete information, then they can be used to make judgements when uncertainty is present.
- ANNs are capable of incorporating multiple constraints (knowledge of a specific task) and finding optimal combinations of constraints for classification, finding a multi–relational mapping between the input patterns and the target classes.

While ANNs are attractive and are still under active exploration, they have problems. As noted in Section 2.2, ASR tends to advance through identification and clarification of problems, and then through a process of finding solutions to problems or methods for circumventing them.

Some problems with ANNs are that it is not clear (i) how to deal with the time sequential nature of speech; (ii) how to model time–shift invariance; or (iii) how to integrate ANNs into large systems, or to integrate them with other techniques in speech recognition (Bourlard *et al.*, 1992; Waibel and Lee, 1989). MLPs are the most popular type of ANNs and have attracted great attention in attempts to solve the problems mentioned above. MLPs are also directly related to this thesis study and the remaining discussions therefore concentrate on MLPs.

Attempts to attack the sequential processing problem have resulted in different strategies. The most immediate solution has been the use of a fixed time approach in which reference patterns which concatenate as many frames as possible are used, and words that are shorter are padded with silence (Peeling and Moore, 1988). This is used only in word recognition.

Another approach is the *Contextual MLP*. Instead of requiring input to the system at time t of patterns associated with t, it also includes patterns from *t-1, t-2, ...* and *t+1, t+2, ...* as left and right contexts of time t (Bourlard and Wellekens, 1987).

Another solution is the *Temporal Flow Model* proposed by Watrous *et al.* (1987). The model uses a feed-forward network with recurrent links connecting hidden units.

Waibel *et al.* (1987) proposed *Time–Delay Neural Networks* (TDNNs). The basic model of a TDNN is based on an MLP, with each processing unit (node) connected to each node at one level higher by n links representing time t, t-t_1, t-t_2, etc., where t_x represents different delay times.

Recurrent networks have been proposed by Elman (1988), Jordan (1986), Mozer (1988) and Robinson (1989) for sequential processing in speech recognition.

The models have connections with delays and feedback links, providing dynamic and implicit memory.

The discrimination ability of ANNs has already been proved to be useful for small systems with a small vocabulary. How best to use ANNs for large systems with a large vocabulary is still an open question. Some researchers have proposed a hybrid approach. For example the approach of using MLPs to generate probabilities used in HMM systems has been proposed by Bourlard and Wellekens (1990), Bourlard and Morgan (1990), and Morgan and Bourlard (1990).

A different technique combines Learning Vector Quantisation (LVQ) (Kohonen *et al.*, 1988) and HMM. This combines the high discriminatory power of LVQ with the HMM's power to model the sequential nature of speech. The code book used in HMM is generated by LVQ, instead of the conventional k-means code book. Bourlard *et al.* (1992) give a detailed coverage of the development of other hybrid models.

In summary, the connectionist approach has been a research area of many years, and has recently received renewed interest. Many problems related to the ANNs themselves or to their applications are being investigated, and elucidated if not solved.

2.3.5 Knowledge–Based Approach

The *knowledge–based* approach became very popular when the *expert system* technique was proposed in artificial intelligence. The basic idea of this approach is to separate the knowledge that is used in the decision making process from the strategies that are adopted to make decisions. A systematic view of this approach is expressed in Figure 2.3. The *knowledge base* contains rules for manipulating the data. These rules are obtained from human experts via the human–machine interface. The control structure (*inference engine*) decides how best to use the knowledge available in the knowledge base and to make decisions. Knowledge is normally explicitly expressed as *rules* in the format of:

if facts *then* Conclusion1 *else* Conclusion2.

The *Goals* are expressed in the conclusions. There are two main techniques for the *inference engine* (decision maker). One is the *Forward Chaining or Data Driven* technique, where the decision maker takes the input, finds applicable rules and fires them until a goal is obtained. Another technique is *Backward Chaining or Knowledge Driven* where the *inference engine* tries to match the goals to the input by applying rules in the knowledge–base and checking the firing result against the input.

A knowledge–based approach has intuitive appeal in ASR. Over the last sixty years, a significant base of speech related knowledge has accumulated. Such knowledge has been reported widely in the literature, including the following forms: Acoustic (Fant, 1960; Flanagan, 1972), Acoustic–Phonetic (Fant, 1973; Lehiste, 1967; Zue and Lamel, 1986), Phonological (Oshika and Zue, 1975; Zue, 1985), Prosodic (Lea, 1980), Syntactic (Woods, 1974), Semantic (Barr and Feigenbaum, 1982).

Two techniques for applying a knowledge–based approach to speech recognition can readily be identified. One is a pure knowledge engineering approach, which incorporates explicit expert knowledge directly. Acoustic–phonetic knowledge is usually derived from careful study of spectrograms and is incorporated using rules and procedures. The facts (acoustic evidence), the knowledge and the strategies are normally extracted from the human expert when reading spectrograms. Examples of this kind of system are: Stern *et al.* (1986), who aimed to study a specific set of phonemes for a specific speaker; Zue and Lamel (1986) who studied a set of phonemes in a specific context for any speaker; and Cole *et al.* (1986) who studied speaker independent recognition of a 'spoken alphabet' in isolated utterance form.

The problem with the pure knowledge engineering technique is that it is difficult to represent the human expert decision rules symbolically, because some components of the rules rely on the use of subtle visual cues by the expert.

A second technique within the general thrust of a knowledge–based approach is indirect use of knowledge to guide the design of models and algorithms based on other techniques, such as template matching and stochastic methods. One example is in work conducted at MIT where the speech signal is segmented and labelled by using a knowledge–based approach. The output of the segmentation process is *dendrograms* which segment the signal into different levels from fine to coarse (Zue, 1988; Glass, 1988).

A pure (direct) knowledge engineering approach has been proved to be manageable for small vocabularies, but it becomes intractable when the vocabulary size increases, that is when it has to manage acoustic–phonetic, lexical, syntactic, semantic, and prosodic facts, and interactions among them. Mechanisms to cope with large a vocabulary and multiple knowledge sources have been investigated by several researchers. Mercier *et al.* (1989) proposed a modular approach which combines speaker independent segmentation with speaker dependent phonetic discrimination to create word lattices that are parsed according to a context free grammar. Another example of a system which copes with a variety of knowledge sources is the Hearsay–II system reported by Erman and Lesser (1980), which incorporates the novel "blackboard" idea to handle multiple and parallel sources of interdependent knowledge. Each source receives input and fires according to information in the blackboard, and write output back to the blackboard. A successor system, Harpy uses a search–based algorithm for recognition (Lowerre and Reddy, 1986) in order to overcome the "bottleneck" problem which occurs with the blackboard technique.

In summary, the extensive literature on speech and related areas provides great insights into understanding of human speech processing, production and perception. A system in which such knowledge is explicitly incorporated simplifies understanding of the performance of the system, and in turn makes error analysis and system enhancement easier. Practical experience with the operation of such systems has great potential to refine our understanding of human speech

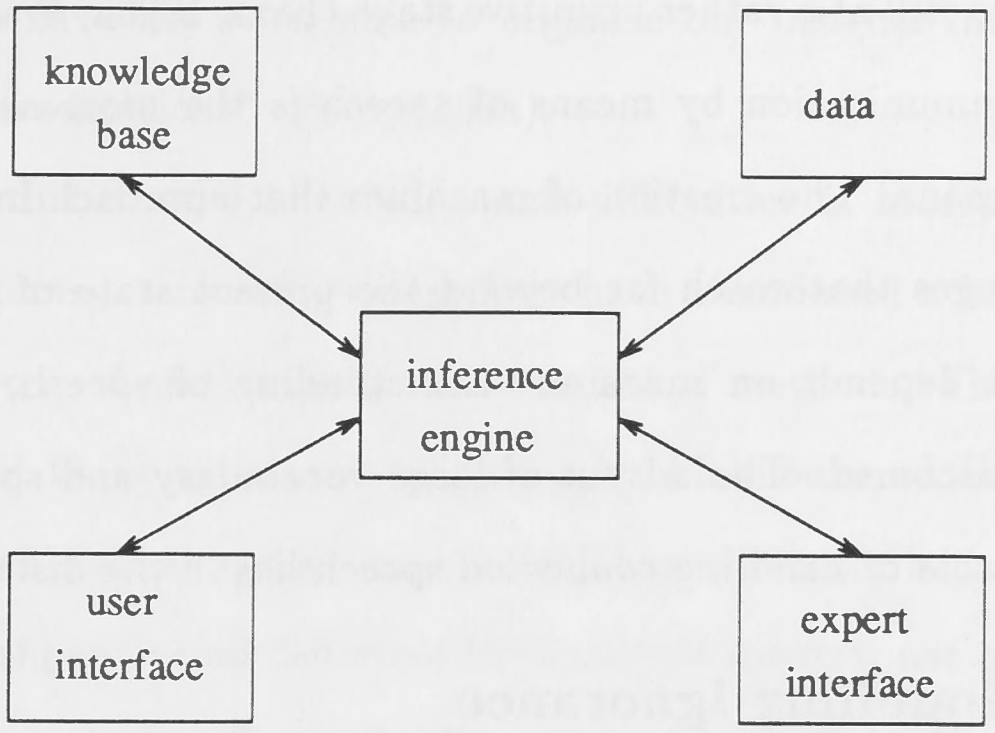


Figure 2.3: Structure of an expert system (Ainsworth, 1988, p. 80).

processing. The most obvious shortcoming of the knowledge–based approach is the difficulty of adequately quantifying human expert knowledge in terms of rules. A further difficulty is that of obtaining sufficiently detailed expert knowledge relevant to large vocabulary continuous speech recognition.

2.3.6 Summary

Sections 2.3.2 through 2.3.5 have discussed four schools of thought in speech recognition, where three of them (the template matching, stochastic and connectionist approaches) can be classified as pattern recognition techniques. Pattern recognition techniques are mathematically or statistically based models which incorporate only limited speech knowledge. These models are intended to model spectral variations as well as temporal variations in the speech signal, with a minimum of human intervention, and eventually to achieve satisfactory recognition. Section 2.4 considers whether or not pattern recognition approaches form a suitable basis for future work.

2.4 The Way Forward

2.4.1 Challenges in Automatic Speech Recognition

The preceding sections described the substantial research effort to date and the limited but encouraging progress. We must admit however that automatic speech

recognition is still at a rather primitive stage (Fant, 1990a; Waibel and Lee, 1990, p. 2). Communication by means of speech is the most natural and efficient form for humans. The creation of machines that approach human ability poses great challenges that reach far beyond the present state of the art, as such a development depends on machine understanding of speech, language and the domain of discourse. The advent of large vocabulary and speaker–independent systems capable of handling connected speech lies in the distant future.

2.4.2 Modelling Ignorance

ASR is clearly a difficult task. The speech signal can vary in many dimensions, largely as a result of speaker-specific factors, ranging from physiological to socio-linguistic. Variations can also appear as a consequence of contextual influences. More specifically, several aspects emerge which can have a significant impact on recognition performance and/or ease of design.

- The nature of the speech — isolated, connected or continuous;
- The size of vocabulary;
- The extent of constraints present, such as task and language constraints;
- The variability among speakers;
- The extent of acoustic ambiguity and/or confusibility;
- The magnitude of environmental noise.

Great effort has been devoted to understanding the human speech communication process and a substantial body of speech knowledge has been accumulated. One can observe, however, that although this knowledge is solid it is far from complete. Available speech recognition technology rests on this limited understanding of the human speech communication process (Levinson and Roe, 1990). "We may have acquired a reasonable overall qualitative view of the nature of

speech, but we have not been able to organise our insights into quantitative, operational representations" (Fant, 1990a).

Because of this limited knowledge and difficulties of incorporating it into recognition systems, many ASR researchers have abandoned a knowledge–based approach in favour of mathematical and statistical methods. This kind of modelling is referred to as *ignorance modelling* by Makhoul and Schwartz (1986). Examples of such modelling are: use of *Euclidean distances* because of a lack of understanding of perceptual distances between two spectra; use of *dynamic time warping* to deal with the effects of different time bases for articulation; and use of statistical modelling (such as hidden Markov models) for speech recognition because of our lack of understanding of how speech signal vary in time.

2.4.3 Overcoming Ignorance: the Need for a Knowledge-Based Approach

On one hand, there is great challenge in making computers understand speech, hopefully as well as human beings. On the other hand, there is still much speech knowledge to be uncovered and difficulties to be overcome in incorporating existing knowledge in speech recognition systems. As Fant (1990a) asked: "What strategy should we choose? Can we pursue a knowledge–based approach and eventually break the speech code, or shall we attempt to train our computers to learn the task by statistical inference? Shall we leave it to the computers to handle a problem that we have failed to formulate and structure in working code?" The ignorance modelling approach is a shortcut taken by many researchers. But "Without a sound knowledge approach we become the servants of computers instead of their masters" (Fant, 1990b).

Fant argues that with better models at all levels of the speech chain (including language theory, speech production and speech perception), we should be able to do a better job in speech synthesis and speech recognition (Fant, 1990a, 1990b).

Zue (1985) emphaseses that abandonment of a phonetically based approach in favour of general pattern matching techniques is partially motivated by our

inability to reliably extract phonetic information from the speech signal. This in turn has slowed advances in acoustic phonetics and phonology, as well as adding to confusion about which form of signal representation is most appropriate for speech recognition. Zue further highlights the importance of the utilisation of relevant available knowledge in speech recognition and of finding procedures that enable us to apply new knowledge as it becomes available.

Zue further argues that speech is a special kind of signal. It is produced with constraints imposed by the language, and by human production and perception mechanisms. Future success in speech recognition will rely heavily on our ability to properly incorporate such knowledge into recognition systems. Zue suggests that speech knowledge can be incorporated at two levels: the first level is to discover those acoustic attributes useful for phonetic contrasts and to determine reliable algorithms to automatically extract these attributes. These are problems in the areas of signal processing and feature extraction. The second level is how to combine these acoustic attributes in order to make phonetic distinctions. These are problems of knowledge representation and control strategy.

Huckvale (1990) argues that the definition of speech recognition problems implies that prior knowledge of linguistic analysis is essential for its solution and suggests that poor current exploitation of such knowledge is a consequence of contemporary pattern recognition architectures. Huckvale argues that there exist two levels of knowledge: acoustic–phonetic, and phonological. The problem with our current ASR systems is that we lack architectures which can exploit such knowledge. He further argues that a network architecture for the lexicon provides a mechanism for the incorporation and exploitation of a range of phonological analyses.

2.5 Concluding Remarks

As argued in the preceding sections, on the one hand it is important to pursue a knowledge–incorporating approach for speech recognition, in order to gain

more insights regarding speech knowledge. Eventually, we will overcome our ignorance by having a full understanding of the speech code. On the other hand, a pure knowledge–based approach needs a complete specification of speech related knowledge in a knowledge base in the form of explicit rules to assist the speech recognition decision process. Our understanding of speech is very limited at present; methods for the optimal representation and effective use of such knowledge still remain in question.

In order to handle this dilemma, this study takes an approach in which a knowledge–based approach is combined with pattern recognition techniques. Making use of ignorance models in order to make best use of acquired knowledge has been suggested by Zue (1985). Fant (1990b) also argues that we sometimes need to use pattern recognition techniques as a tool in the quest to gain a deeper understanding.

There are two levels at which knowledge can be included in an ASR system. That is at the acoustic–phonetic level and at the phonological level. This study investigates ways of incorporating knowledge at the acoustic–phonetic level. Unlike traditional knowledge–based approaches, which incorporate knowledge from spectrogram reading, this study explores techniques for incorporating linguistic–phonetic knowledge, i.e. phonetic features.

The acoustic signal clearly includes not only linguistic information, but also non-linguistic information, such as speaker identity, the speaker's physiological and psychological states, and the acoustic environment. What primarily concerns an ASR machine at present is the linguistic information, the non-linguistic information being important only to the extent that it allows deconvolution of such "extraneous" effects from the linguistic information. By encoding distinctive features, we expect that linguistic information can be extracted, and that less emphasis will be placed on the non-linguistic information, thereby achieving a degree of normalisation.

Traditional ASR systems are based on assumptions about the acoustic properties of the speech signal that are useful for discriminating between words in

the vocabulary. Although these acoustic properties include all possible variations as well as linguistic information, general pattern recognition techniques directly compare the acoustic properties of unknown signals with the acoustic properties of known reference signals. This kind of system will thus have difficulty coping with certain types of variability, as the acoustic properties change according to speaker, speaker's emotional status, etc. With an early transformation of acoustic data to a phonetic domain, we expect that essential acoustic properties of the linguistically relevant information can be extracted, with recognition then being performed in the phonetic feature domain. We expect that under this paradigm, the methodology will move forward in simulating human speech recognition processes. This paradigm should also simplify understanding of the system's behaviour, facilitating analysis of errors at different levels of the system, and subsequent modifications to the system. It builds an environment in which speech related questions can be investigated, such as acoustic properties of the phonetic features, binary versus continuous properties of the phonetic features, existence of invariant cues in the speech signal, etc. It should also allow us to gain further insights into the speech signal. These and related topics are investigated in the following chapters.

Chapter 3

Distinctive Features, Invariance, and Sub-Phonetic Units: A Review

3.1 Introduction

The last chapter introduced one of the major aims of this study, that is incorporation of relevant acoustic-phonetic knowledge in a speech recognition system. This approach simplifies conceptualisation of the system, makes analysis of the performance and error analysis less complicated, and facilitate subsequent improvement to the system. Success in this aim in turn helps enhance the performance of the ASR system and our understanding of the speech signal. The form of knowledge that we have chosen to integrate is that of *distinctive features.* In this chapter, we describe distinctive features in detail by introducing different theories of distinctive features, analysing the foundations and role of the distinctive features, and reviewing different opinions about the nature of distinctive features — in particular, are they continuous or binary. A central question arises: Is there a set of acoustic correlates for each distinctive feature? This question is directly related to the lively debate concerning the existence of invariant cues. If we can discover such sets of acoustic properties which characterise distinctive features, this will provide strong evidence for the existence of invariances in the speech signal. After describing distinctive features, this chapter moves on to analysis of the literature concerning these questions. The final section reviews research efforts concerning practical systems that make use of distinctive features for speech processing and recognition. This review addresses the questions posed above from a practical

point of view. In summary, this chapter establishes the theoretical framework of the thesis.

3.2 Distinctive Feature Theory

3.2.1 A Historical View of Distinctive Feature Theory

Jakobson and Waugh (1979) date the history of distinctive features as starting in 1928, when the preparatory committee of the first International Congress of Linguists raised the question of methods appropriate to a comprehensive view of a given language. In response, the Prague School outlined a particular class of significant differences, namely a phonological correlation "constituted by a set of binary oppositions all of which are defined by a common criterion conceivable apart from each couple of opposites." The phonologist Nikolaj Sergevič Trubetzkoy was an important contributor to this view.

After intensive discussions on the formulations of phonology at two of the aforementioned congresses, and subsequently in a more comprehensive way at the Prague International Phonological Conference of 1930, the need for a consistent definition of a phoneme in terms of its simultaneous components was made increasingly clear. In the early 1930s, the Prague School designated the phoneme as "a set, bundle, totality of these concurrent sound properties which are used in a given language to distinguish words of unlike meaning." These properties were initially designated as different or distinctive qualities or attributes (Jakobson and Waugh, 1979). They were later referred to by Bloomfield (1933) as *distinctive features.*

Trubetzkoy (1939) describes his phonological investigations in relation to clarification of the nature and status of distinctive features. The original paper in German was translated into English by Baltaxe (1969).

Distinctive feature theory has gone through continuous development, as summarised by Jakobson and Waugh (1979) and Baltaxe (1978).

3.2.2 Foundations of Distinctive Features

Speech sounds are the product of articulation of the human speech apparatus. Speech signals are the acoustic representation of these sounds. Different articulations yield different speech signals. The signals may be characterised in terms of intensity, duration, energy, voice fundamental frequency F_0, formant structure, etc. Clearly, speech sounds possess complexes of properties, and these properties are shared across of the range of speech sounds (Jakobson, 1938; Jakobson *et al.*, 1952). Stevens (1980) analysed certain basic acoustic properties that help to distinguish classes of speech sounds from each other. On the basis of his analysis, he argued that an essential part of the speech-processing capabilities of a language user is a classification of speech sounds in terms of these basic properties, i.e. features.

To define the term "feature", i.e. *distinctive feature*, another concept has to be introduced: that of *minimal distinction*. A distinction is called *minimal* if it cannot be divided into further distinctions that are used to differentiate words in a given language (Jakobson *et al.*, 1963). For example, the distinction between the word *bill* and *pill* is minimal, as a result of the distinctive feature *voicing*: [**b**] being voiced and [**p**] unvoiced. The distinction between *pill* and *till* is minimal too: the distinction between [**p**] and [**t**] is in place of articulation, represented by the feature pair *grave/acute*. The distinction between *pill* and *dill* is not minimal, as the distinction consists of [**p**] versus [**t**] with respect to place of articulation (represented by the feature pair *grave/acute*), and [**t**] versus [**d**] with respect to *voicing*.

Features (e.g. *grave, voicing*) which serve to distinguish between sounds normally have two values (e.g. *acute/grave*), the listener being obliged to choose one of them, such as *acute* versus *grave*, or between presence and absence of certain qualities. *Distinctive features* have this property. They are ultimate distinctive entities of language which cannot be broken into smaller linguistic units. Within

the *distinctive feature* context, a phoneme is formed by a bundle of concurrent distinctive features (Jakobson *et al.*, 1963).

Fant (1986) states that the two main purposes of distinctive features are to develop a language universal system of phonetic categories (meaning they can be used to make unique distinctions between phonemes in all languages) and to describe essentials of the speech code (by relating distinctive features to acoustic properties). Combining these two purposes, and relating them to the speech communication process, Fant suggests that phonemes are coded by distinctive features. Distinctive features are both physical (relating to the domain of articulation and acoustics) and psychological (relating to the perceptual domain) realities of phonemes (Singh, 1976; Fant, 1966, 1973). The relationship of one phoneme to another can be formulated via distinctive features (Baltaxe, 1978).

3.2.3 Different Theories Regarding Distinctive Features

Distinctive feature theory has evolved over the 60 years since it was first set forth. Many researchers have contributed and are still contributing to the field. Within the *distinctive feature* framework, many theories have also been developed, such as those which are perceptually based, articulatorily based and acoustically based. Although there are numerous descriptive systems extant, there are basically only two ways to derive the features (Fant, 1969, 1973):

- Select an inventory of classes suitable for encoding language structures, then determine their phonetic correlates; or
- Make an analysis of the modes and constraints of the speech production mechanisms and associated perceptions and determine their distinctive function in language.

Fant (1969, 1973) points out that distinctive feature theory has developed along both those lines. Investigators differ only in the relative importance placed on one or other method.

The following sub-sections aim to briefly describe the major distinctive feature theories. The interested reader can pursue these theories in greater detail by consulting the referenced original papers.

3.2.3.1 Trubetzkoy's Theory

Trubetzkoy, one of the founders of the Prague School of Linguists, developed his *Principles of Phonology* (1939) in the decade preceding World War II. He classified distinctive oppositions[1] on the basis of

- the relationship between opposition members,
- their relationship to the entire system of oppositions,
- nature of oppositions, and
- the extent of their distinctive force.

Trubetzkoy attempted to perform a phonological analysis of phonetic contrast (Hyman, 1975), but his theory is very complex (Baltaxe, 1978). The remainder of this sub-section briefly introduces the main characteristics of Trubetzkoy's work.

Relationship Between Opposition Members

Trubetzkoy (1939, 1969) differentiated the relationship between opposition members by the terms bilateral and multilateral. Bilateral means that only two opposition members share the sum of phonetic features. For example, in English, /**p**/ and /**b**/ are two members of a distinctive opposition and have in common some features: they are both "oral labial stops". The opposition is bilateral since there are no other consonants in English which come under the heading "oral labial stops".

Multilateral means that more than two members share the common phonetic features. For example, in English, /**f**/ and /**b**/ are two consonants which are

[1] opposition means a sound difference that results in a meaning difference.

both "labial obstruents". This is an example of a multilateral opposition, since /p/ and /v/ are also "labial obstruents" in English.

Relationship to the Entire System of Oppositions

There are two types of oppositional relationships: proportional and isolated (Trubetzkoy, 1939, Baltaxe, 1969). Trubetzkoy states in translation (Baltaxe, 1969, p. 70) that "an opposition is proportional if the relationship between its members is identical with the relation between the members of another opposition or several other oppositions of the same system". Otherwise the opposition is said to be isolated.

Nature of Oppositions

When determining the nature of an opposition, it is always important to consider the inventory of distinctive sounds (phonemes) in the language under investigation. There are three types of oppositions used for classifying oppositions on the basis of the relation between the members of an opposition:

- Private (Trubetzkoy, 1939, p. 75)

  In a bilateral type of opposition, one member of the opposition carries a phonetic "mark" (presence) which the other member lacks (absence). The opposition member which has the mark present is "marked", while the other member is "unmarked". For example, in the opposition /**b**/ : /**p**/ in English, /**b**/ is characterised by the presence of *voicing*, while /**p**/ lacks *voicing*.

- Gradual (Trubetzkoy, 1939, p. 75)

  In gradual oppositions, the members are characterised by different degrees of the same property. One example is vowel height in English, which can be used to distinguish front vowels where there is a continuum of vowel height.

- Equipollent ("logically equivalent") (Trubetzkoy, 1939, p. 75)

  An opposition is said to be of type "logically equivalent" when it is not possible to view the two members as differing by the degree of some phonetic

property, nor at the same time is it possible to view one as having a mark which the other lacks. Consider for example, in English, /**p**/ and /**t**/, or /**t**/ and /**k**/, the different places of articulation are obtained by discrete changes in the two articulators (upper and lower lips for /**p**/; tip of the tongue and the alveolar ridge for /**t**/) as with most consonants.

Extent of the Distinctiveness

Trubetzkoy (1939) made another classification according to the extent of the distinctiveness of an opposition (p. 77) which is *constant* opposition versus *neutralisable* opposition.

When one member of an opposition always occurs in a certain position phonetically (e.g. at the end of a word), the opposition is said to be neutralised.

When two members of an opposition can occur in all positions, the opposition is said to be constant.

This concludes the brief review about Trubetzkoy's (1939) distinctive feature theory. Interested readers can refer to the original book or to the translation into English by Baltaxe (1969).

3.2.3.2 Jakobson, Fant and Halle's Theory

Jakobson — a friend of Trubetzkoy (Baltaxe, 1978) — together with his colleagues Fant and Halle developed their theory of *distinctive features* (Jakobson *et al.*, 1952, 1961). Their motivation was to develop a theory of phonology that would predict the only oppositions that could be found in languages. Simplicity, economy, and generality were among the primary aims of their work, whose underlying basis was the description of the system of sounds. Their departure point was that speech sounds pose complexes of properties, and these properties are shared across groups of speakers. The evidence was presented in acoustic terms, mostly utilising spectrograms. Jakobson *et al.* were able to present clear distinctions between minimally distinct pairs of phonemes using spectrograms. After

examining the systematic acoustic properties of the sound system, they related these acoustic findings to articulation and to the perceptual domain. They examined a number of languages and developed the strategy of breaking the phoneme into its ultimate units, i.e. distinctive features. The distinctive features were considered to be the ultimate units because they cannot be resolved into finer units of distinction. Jakobson and Waugh (1979) emphasised the importance of an acoustic basis and insisted that speech communication is achieved by the nature of sounds and not by the nature of their production. They claim (p. 40) that all sounds in the languages of the world can be distinguished using 12 pairs of binary distinctive features:

1. Vocalic/Non-Vocalic;

2. Consonantal/Non-Consonantal;

3. Interrupted/Continuant;

4. Checked/Unchecked;

5. Strident/Mellow;

6. Voiced/Unvoiced;

7. Compact/Diffuse;

8. Grave/Acute;

9. Flat/Plain;

10. Sharp/Plain;

11. Tense/Lax;

12. Nasal/Oral.

They used 9 of their 12 features to describe 28 phonemes of Received Pronunciation English (Jakobson, 1961, p. 43).

Relationship to Articulation

Jakobson *et al.* (1952) used two major classes of features (*vocalic* and *consonantal*) to distinguish four major classes of segments in the following way:

True consonants:

$$\begin{pmatrix} +cons \\ -voc \end{pmatrix}$$

Vowel:

$$\begin{pmatrix} -cons \\ +voc \end{pmatrix}$$

Liquid:

$$\begin{pmatrix} +cons \\ +voc \end{pmatrix}$$

Glide:

$$\begin{pmatrix} -cons \\ -voc \end{pmatrix}$$

Vowels

The phonetic properties of tongue height, tongue position and lip rounding of the vowels are characterised by four features in Jakobson *et al.* (1952): *diffuse, compact, grave* and *flat*. The correlation of these features with the articulatory description of vowels is summarised in Table 3.1.

Consonants

The consonants are characterised by six features in Jakobson *et al.*'s feature system. These are *diffuse, grave, voice, continuant, strident* and *nasal*, whose correlation with the consonants is summarised in Table 3.2.

Summary

The characteristics of Jakobson *et al.*'s distinctive feature system are that:

- it relates distinctive features to distinct sounds on the basis of their acoustic properties, differing from traditional phonetic features which are more articulatorily based;

Features	*Vowels*
+diffuse	high vowel
–diffuse	mid and low vowel
+compact	low vowels
–compact	high and mid vowels
+grave	back vowels
–grave	front vowels
+flat	rounded vowels
–flat	unrounded vowels

Table 3.1: Correlation of Jakobson *et al.*'s distinctive features with the articulatory description of vowels (Hyman, 1975, p. 36).

Features	*Consonants*
+diffuse	labial and dental/alveolar consonants
–diffuse	palatal and velar/back consonants
+grave	labial and velar/back consonants
–grave	dental/alveolar and palatal consonants
+voice	voiced consonants
–voice	voiceless consonants
+continuant	fricatives, liquids, glides
–continuant	stops and affricates
+strident	noisy fricatives (labiodental, alveolar, alveopalatal), affricates
–strident	less noisy fricatives (interdental, as well as palatal and velar), stops, liquids, glides
+nasal	nasal consonants
–nasal	oral consonants

Table 3.2: Correlation of Jakobson *et al.*'s distinctive features with the articulatory description of consonants (Hyman, 1975, p. 39).

- it relates these acoustically based distinctive features to articulation and perception;
- it classifies consonants and vowels in the same system of distinctive features, whereas some other distinctive feature systems classify the vowels and consonants separately (e.g. Chomsky and Halle (1968));
- the features are all binary in nature;
- the distinctive feature system is compact, as one of the aims of their system was to have it as economic as possible, i.e. to use a minimal set of features.

While a combination of acoustic, articulatory, and perceptual aspects in the definition of distinctive features was seen as most welcome by some researchers (Fant, 1966, 1973), others had reservations about this theory. Fant (1973) comments that their distinctive features are not intended as absolute descriptors of the spectrographic qualities, in that the articulatory and the acoustic properties of any feature will vary a little with different contexts (phonetic and feature). Therefore, the invariance is generally relative rather than absolute.

While on one hand Jakobson *et al.*'s theory is compact and economic, on the other hand, use of the same features to describe both consonants and vowels is somewhat clumsy because of their different articulation; some features applicable to vowels are not logically applicable to consonants (Fant, 1966, 1973). Other researchers, however, saw advantages in this uniformity of treatment (e.g. Hyman, 1975).

The use of binary distinctive features seems to be simple and uniform, but some sounds cannot be distinguished by use of only binary features. One example is the inability of distinguish different vowel heights, (Brakel, 1983, p. 54), as there are more than two vowel heights in English.

Another limitation of Jakobson *et al.*'s distinctive feature theory is that it lacks a realistic discussion of time varying aspects of speech patterns (Fant, 1966, 1973).

3.2.3.3 Chomsky and Halle's Distinctive Features

The distinctive features presented in Chapter VII of *The Sound Pattern of English* (Chomsky and Halle, 1968) are designed on one hand to capture the phonological contrast of languages in the same way as Jakobson *et al.*'s features. On the other hand, they are designed to describe the *phonetic* contents of segments derived by phonological rules. Chomsky and Halle's features are based on the phonetic and articulatory possibilities of the human vocal apparatus.

Chomsky and Halle described the articulatory features of the universal sounds on the assumption that the configuration of the human vocal mechanism and the speech perception mechanism are identical for all humans. Under this assumption, the vocal mechanism was considered in terms of the source of sounds, the different areas of the vocal tract involved, and the different positions of the tongue in relation to different areas of the vocal tract. Each feature is binary. If Trubetzkoy's features are regarded as the most complex ones, and Jakobson *et al.*'s as the simplest, then Chomsky and Halle's features are intermediate.

There are five major categories in Chomsky and Halle's feature system:

- major class features;
- cavity features;
- manner of articulation features;
- source features; and
- prosodic features.

Within the major class feature category, there are three pairs of features:

- consonantal/nonconsonantal;
- vocalic/nonvocalic,
- sonorant/nonsonorant.

These pairs are used to divide broad phonetic categories into consonants, vowels and vowel-like consonant sounds.

Within the cavity group, there are eight types of features:

- coronal/noncoronal;
- anterior/nonanterior;
- tongue body features: high/nonhigh, low/nonlow, back/nonback;
- rounded/unrounded;
- distributed/undistributed (concentrated);
- covered/uncovered;
- glottal constriction and secondary apertures (nasal/nonnasal, lateral/non-lateral).

Within the manner of articulation feature category, there are four types of features:

- continuant/noncontinuant;
- release features: instantaneous release/delayed release;
- supplementary movements: suction, pressure and order of release in sounds with multiple closures;
- tense/nontense.

Within the source feature group there are three types of features:

- heightened subglottal pressure;
- voiced/unvoiced (voiceless);
- strident/nonstrident.

Chomsky and Halle (1968) do not present any final conclusions regarding prosodic features, simply referring to Wang (1967).

Chomsky and Halle's feature theory also has been subject to critical comments. Fant (1969, 1973) criticised Chomsky and Halle's purely speech production categorisation approach. He pointed out that features were not constrained by the phonetic components of speech events and nor related to psychological reality. Fant emphasised the importance of relating a feature system to the speech communication chain: that is, to relate the feature system to acoustic and perceptual correlates and not to stop just at the production level.

Others have more specific concerns. Singh (1976) commented from a practical point of view that Chomsky and Halle's feature system may be satisfactory theoretically, but if it is unable to account for speech production and speech perception errors, it does not contribute to the understanding and treatment of phonological problems of deviant speakers.

Brakel (1983) demonstrates that the descriptions proposed in *The Sound Pattern of English* are relatively inadequate as phonetic primes (features) for phonological description. He aimed to reduce the number of primes and strengthen the basic hypothesis[2], at the same time making the features phonologically and physiologically more accurate. He modified Chomsky and Halle's feature system to some extent.

Reenen (1982) analysed *The Sound Pattern of English*, modifying and extending the framework presented therein, especially with regard to the feature *nasal*.

3.2.3.4 Hierarchical Feature Systems

While all the feature systems summarised above were based on a bundle of features acting in parallel, authors Singh and Singh (1976), Sagey (1990), and McCarthy (1988) suggest that the features should be organised hierarchically.

[2]A distinctive feature inventory is a hypothesis about all the phonological systems in the world. It says that these and no more than these features are necessary to characterise all the phonological systems employed by humans.

Singh (1976) tried to reorganise the existing feature systems such that sounds could be arranged in phonetic classes by going through a tree-like structure. The division of groups of features is finer and finer from the root of the tree to the leaves of the tree. For example, he presents a hierarchical organisation for Jakobson *et al.*'s feature system (Singh, 1976, p. 40).

In contrast, other researchers have proposed a hierarchical organisation of features into functionally related classes, grouped under nodes of a tree structure. Their aim is to try to solve a fundamental problem in phonological theory, namely, that processes often operate on a consistent subset of distinctive features within a segment, such as the set for characterising the place of articulation. McCarthy (1988) gives a good review of this approach and presents evidence for the existence of a feature hierarchy.

3.2.3.5 Other Distinctive Feature Systems

The previous sections have concentrated on the feature systems of Trubetzkoy, of Jakobson *et al.*, and of Chomsky and Halle. This section briefly describes other systems, further details of which can be found in the original publications. Singh (1976) also gives a brief introduction and in addition compares several systems for consonants.

Apart from Jakobson *et al.* (1952) and Chomsky and Halle (1968), other systems concentrate either on vowels or on consonants separately, because their authors considered that consonants and vowels have distinctly different systems of articulation and hence different systems of acoustic and perceptual correlates.

Vowels

Vowels are less controversial than consonants; most authors agree on their basic properties. A common aspect of the feature systems referred to here is that the hypothesis tested was that the known acoustic and/or articulatory parameters of vowels may be utilised as perceptual cues. Specifically, these systems are perceptual feature systems for vowels — Shepard (1972), Hanson (1967), Singh

and Woods (1971), Terbeek and Harshman (1971), and Anglin (1971). All of these are perceptual feature systems for English vowels, except that Hanson's (1967) is a perceptual feature system for Swedish vowels. The perceptual feature systems are derived by analysing the experimental results from listeners' judgements of sounds.

Consonants

Feature systems for consonants are the subject of more controversy than those for vowels, the basic disagreement being on the definition of phonemes (Singh, 1976). Some systems define a phoneme on the basis of acoustic evidence (e.g. Jakobson *et al.* (1952)), some on the basis of articulatory evidence (e.g. Chomsky and Halle (1968) and Halle (1964)), some on perceptual evidence (e.g. Wickelgren (1966) and Singh, Woods and Becker (1972)) and others on a combination of articulatory and acoustic evidence (e.g. Miller and Nicely (1955), and Singh and Black (1966)).

As noted earlier, Singh (1976) gives a systematic comparison of the above feature systems in the context of consonants.

3.2.3.6 Discussion

Most of the distinctive feature theories reviewed in this section regard the *phoneme* as a bundle of concurrent distinctive features. Although the initial motivation of some distinctive feature theories was to investigate how a real speech sound is recognised by a listener (e.g. Jakobson *et al.* (1961)), the phonetic variations of a phoneme in speech sounds are hardly addressed in the context of distinctive feature theories. A phoneme can be pronounced differently when it is in a different phonetic context. Ladefoged (1982, pp. 82–88) summarised some rules for allophonic realisation in English to describe the phonetic variations of phonemes. Three of those examples are stated here: (i) obstruents — the stops and fricatives — classified as voiced are not in fact voiced throughout the articulation when they occur at the end of an utterance or before a voiceless sound (e.g. /**v**/ in "try to

improve" or /**d**/ in "add two"); (ii) vowels tend to become nasalised before nasal consonants (e.g. "ban") because of anticipatory coarticulation effects; and (iii) all the front vowels become considerably retracted when they occur in syllables closed by /**l**/ (compare the pronunciation of /**i**/ in "heed" and "heel", and of /**æ**/ in "pad" and "pal"). Although a combination of concurrent distinctive features is regarded as a phonemic pattern in the literature, the position taken herein is that this combination describes sounds which are more phonetically based. This phonetically based pattern should be one of the phonetic variations of a given phoneme. Oasa (1980) in a different type of study demonstrated a similar point, where he used phonetic features to analyse some Australian English vowels from different dialect regions. The outcome of this analysis was phonetically based, to demonstrate that in forming the same phoneme from phonetic patterns from different dialect regions a set of rules had to be applied.

3.2.4 Roles of Distinctive Features

The significance of the concept of distinctive features has became apparent through perceptual experiments and linguistic debate (Liberman, 1974). Jakobson and Waugh (1979) discussed the significance of distinctive features in great detail. Jakobson and Waugh (1979) quoted from Balonov and Deglin (1976, p. 182) who claimed from a neuro-linguistic point of view that: "it becomes evident that as a rule the left hemisphere brings about the classification of phonemes on the basis of their distinctive features and supports the hierarchy of these features ensuring the stability of the phonological system of language."

Apart from being a phonological tool which assists in organisation of phonemes, distinctive features are directly related to perception. Fant (1966, 1973) insisted that distinctive features are phonetic classes, and a psychological reality as judged from perceptual experiments. Jakobson and Halle (1956) argued that distinctive features are language specific. For each language, there is a corresponding set of distinctive features; the phonemes of these languages are comprised by bundles

of concurrent features and phonological rules define what phonemes can be concatenated together. If a listener receives a message that he/she knows, he/she correlates it with the code which includes all the distinctive features to be manipulated, all their admissible combinations into bundles of concurrent features (i.e. phonemes) and all the rules for concatenating phonemes into sequences, to differentiate morphemes and whole words. There have been several investigations of relationship between distinctive features and perception reported, such as the examples described below.

Blumstein and Cooper (1972) performed dichotic perception experiments on CV stimuli and reported two main results. First, they demonstrated that single feature contrasts are easier to *identify* than double feature contrasts, due to greater shared information in the case of single distinctive feature difference (the total information which must be extracted and held in order to identify competing stimuli is less). Second, double feature contrasts are easier to *discriminate* than single feature contrasts, due to the implication that additional feature contrasts lead to larger perceptual distances (i.e. to an easier discrimination). Blumstein (1974) reports the results of dichotic perception experiments with CV stimuli: single feature contrasts are easier to *identify* than double feature contrasts. The manner features are generally salient as they separate phonemes belonging to different sound classes (e.g. stops, nasals, fricatives, affricates, etc.). She also demonstrated that "features are independently extracted during the speech perception process" (p.340). Studdert-Kennedy and Schankweiler (1970) in their dichotic listening experiment investigated the perception of CVC' nonsense syllables with competing initial stop consonants. They noted that performance was affected by the particular feature relations held between competing consonants, as well as by the number of feature attributes which the competing consonants shared. Teuber (1976) after experimenting with a patient suffering from certain left hemisphere lesions, and analysing the categorical perception of phonemes in the preverbal child (Eimas *et al.* (1971) showed that babies could readily distinguish /**p** / from /**k** / across different vowel environments), suggests the existence

of "a 'template' for detecting the distinctive features out of which the phonemes of language are built", and concludes that "the 'distinctive features' would be more than a universal schema for classifying phonemes in all their diversity, across language; the features would be 'real', in the sense of being universal neuronal mechanisms for producing and for perceiving sounds of speech" (p. 137).

Singh (1976) put forward distinctive features in a more general framework, arguing that phonemes are interrelated in terms of distinctive features, that these interrelationships of the phonemes are of differing degrees, and finally that these interrelationships reflect on phoneme production, phoneme perception and other processes involved in the encoding and decoding of speech. He further suggested that the sounds of speech are processed in terms of distinctive features, and he emphasised the need to use distinctive feature distances when measuring disparities in speech production and perception. Singh states that when the distance between two phonemes in terms of distinctive features is bigger, there is less chance of making errors between them. The advantage of counting speaking errors in terms of distinctive features is that a more representative indicator of the magnitude and nature of a speaker's problems is obtained than by counting a simple aggregate of incorrect phonemes. The basis of Singh's suggestion is that phonemes of a language are perceived in terms of distinctive features, and that the basic elements of speech sounds produced are distinctive features. He also indicated that when using this distance, some features play a more important role than others, therefore, they should receive more weight.

3.2.5 Distinctive Feature Distances

Singh and Singh (1976) and Singh (1976) proposed that the differences between phonemes should be calculated in the following way: given a table of definitions of phonemes in terms of distinctive features, the differences between two phonemes can be calculated by counting the number of features for which the two phonemes have different feature values.

This distance has been seen as a more meaningful measurement of phonetic distance than the phoneme distance (the number of differing phonemes) itself, taken literally, making it easier to analyse errors made when speaking or when listening in order to reach conclusions about the types of errors and the possible cause of errors (Singh, 1976).

Ladefoged (1971) developed the same idea and extended it to differentiating between languages and/or dialects, with the expectation that on an average basis, such differences will represent their phonetic differences.

Singh (1976) acknowledges that in different distinctive feature systems, the discriminatory power of different distinctive features will vary, and that therefore, they will not differentiate phonemes equally. This point is also made by Fant (1969, 1973).

3.2.6 Continuous versus Binary properties

Trubetzkoy's (1939) distinctive feature system emphasises the continual/gradual nature of features; on the other hand, Jakobson *et al.* (1952), Chomsky and Halle (1968) and other systems have features that are predominantly binary. While Jakobson *et al.*'s binary property had been seen as one of the innovations of their distinctive feature theory, the theory has been criticised for not being able to distinguish more than two degrees of vowel height or of vowel frontness and backness (Brakel, 1983; Hyman, 1975). One explanation given by Hyman (1975) was that Jakobson *et al.*'s aim was to capture phonological oppositions, not different phonetic realisations.

Foley (1970) and Vennemann (1972) argued for phonological features which are represented in variable strength in various consonants and vowels. Ladefoged (1971) proposed multivalued features in phonological rules, i.e. a different degree of glottal stricture and scale based on the state of the glottis.

Fant (1973) commented that in reality features exhibit more than two values and even Jakobson *et al.* admitted four levels of compactness (Jakobson *et*

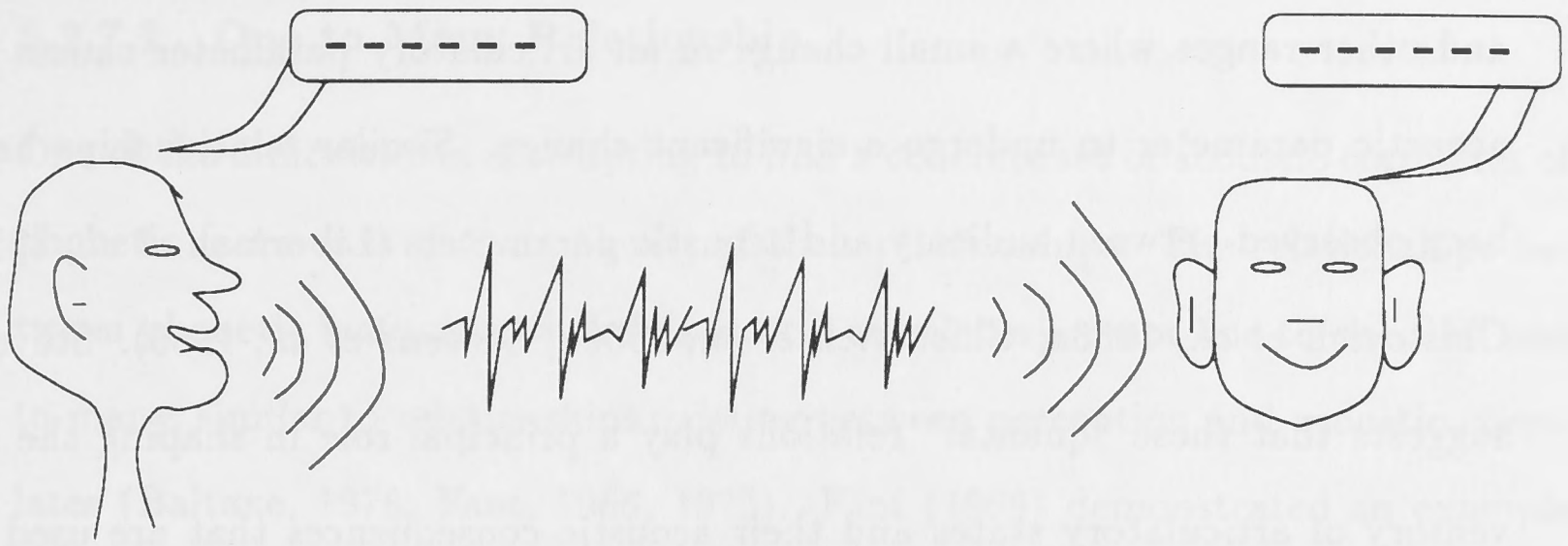


Figure 3.1: Speech communication process.

al., 1961, Section 2.414). Fant (1969, 1973) argued that the binary nature of distinctive features is simply a coding convenience.

Zue and Schwartz (1980) expressed a view that includes wider realities. They suggested that at the phoneme level, distinctive feature theory necessitates a discrete (or even binary) selection, whereas at the articulatory and acoustic levels, the feature correlates appear to take on a continuum of values.

3.2.7 Acoustic Correlates of the Features

In the speech communication process, the speaker communicates to the listener by encoding the intended message into a corresponding set of controlled articulator movements. The result of this set of movements is the speech sound as represented by the acoustic signal. The speech sound is then perceived by the listener, and decoded into meaning (see Figure 3.1). It is natural to think that there is a set of correlates between the articulatory movement and the acoustic signal, and a set of correlates between the acoustic signal and the perception, but attempts to find perceptual and/or acoustic feature correlates to the phonetic features have had only partial success (Wickelgren, 1966; Klatt, 1968; Singh, Woods and Becker, 1972; Wang and Bilger, 1973). Qualitative results regarding these correlates have been reported by some researchers, (e.g. Stevens (1972, 1980), Jakobson *et al.* (1952), Fant (1960)).

Stevens (1972) demonstrated that as articulatory parameters are varied gradually, there are ranges where the acoustic parameters are relatively invariant,

and other ranges where a small change in an articulatory parameter causes an acoustic parameter to undergo a significant change. Similar relationships have been observed between auditory and acoustic parameters (Liberman *et al.*, 1957; Chistovich *et al.*, 1966a; Chistovich *et al.*, 1966b; Stevens *et al.*, 1969). Stevens suggests that these "quantal" relations play a principal role in shaping the inventory of articulatory states and their acoustic consequences that are used to signal distinctions in language. He further suggests that all phonetic features occurring in language may well have their roots in acoustic attributes with these characteristics.

Stevens (1980) demonstrates that the distinction between consonants and vowels is that consonants have rapid changes in their short time spectra and vowels do not. Distinctions between consonants lie in the gross shapes of their spectra, whereas distinctions between vowels depend on the presence of energy in certain frequency regions. Stevens argues that acoustic correlates of phonetic features can be defined in terms of the acoustic attributes of the signal within one of the well defined regions, where the acoustic parameters remain relatively invariant when articulatory parameters vary gradually.

Fant (1990) observes that although speech research has acquired a reasonable overall qualitative view of the nature of speech, it has not yet been able to organise this qualitative view into quantitative and operational representations. Because speech scientists have not yet fully understood relationships between either articulatory configurations and the resulting acoustic signals, or the acoustic signal and acoustic correlates of features (Glass, 1988), "We are plagued with a variability of realisations. We have superficial insight into the overall structure of variabilities, but we have not been able to formulate consistent rules within a wide frame of contexts. We are thus at a loss when attempting to specify invariance criteria" (Fant, 1990a).

3.2.7.1 One to Many Relationship

One of the difficulties in attempting to find a concrete set of acoustic correlates of phonetic features is the complexity of their relationships. The relationships between phonetic features and their acoustic correlates are not one to one, but one to many, similar to relationships existing between perception and acoustic correlates (Baltaxe, 1978; Fant, 1966, 1973). Fant (1968) demonstrated an example of this one to many relationship. He summarised a set of phonetic features and their acoustic correlates, and showed that a single feature is normally specified by multiple acoustic parameters.

3.3 Invariance and Variability

As Fant (1990a), Glass (1988) and several other authors observe, we do not yet fully understand the speech code — the mapping from acoustic signals to articulatory configurations — and there is thus a strong need for further research. Because of our lack of understanding of the speech code, there continues to be a lively debate about the extent to which there are invariant characteristics in the acoustic signal.

As noted earlier, the most universal characteristic of the speech process is its variability. Variability in the acoustic manifestations of a given utterance (which is perceived to have a certain phonetic or phonemic quality) is substantial and arises from many sources, including:

- Intra-speaker variability which includes diverse voice quality, voice fundamental frequency, speaking rate, long term variability, emotional state, stress, physiological condition, etc.;
- Inter-speaker variability which includes differences in vocal-tract anatomy, dialect, accent, detailed articulatory habits, etc.;
- Coarticulation effects which cause the same phoneme to be produced differently in different phonetic contexts;

- Recording conditions which include background noise, room reverberation, microphone/telephone characteristics, etc.

Because of these variabilities, only partial success in speech recognition has been achieved. The immediate question which arises is that of how humans can handle all these variabilities so well. Because the acoustic signal is the medium of speech communication, and the product of articulation, we may hypothesise that there are invariant characteristics in the acoustic signal, some of which may be emphasised to decode the phonetic content which represents the linguistic information and others relating to non-linguistic information can be de-emphasised for some purposes.

There are two schools of thought with regard to the invariance issue (the interested reader can refer to Perkell and Klatt (1986) for a comprehensive review). One school does not believe that invariant cues exist, arguing that the speech signal is highly variable. As one example is, Lindau and Ladefoged (1986) insisted on the variability of the feature specification based on the production of speech, showing the variability of speech production. They point to a lack of invariant cues and claim that there is many to many mapping between phonetic features and associated physical acoustic parameters.

Another example is Pols (1986), who emphaseses contextual effects, and argues that the more context dependent effects we discover when reporting results, the less it can be said that dominant invariant cues exist. Pols also points to the existence of multiple cues contributing to a single phonetic identification. Suomi (1985, 1987) and many others also report that there are context dependent effects in the production of stop–vowel syllables.

The opposing school believes that invariances do exist in the acoustic signal. They attribute disbelief in the invariance theory to the apparent variability of the speech signal, and the difficulty of relating articulation to the acoustic signal, given (i) that any relationship here will be rather complex, and (ii) that our understanding of the speech signal is far from complete (Fant, 1990a). The

difference between a "planned segment" and a "uttered segment" is considerable, the former being envisioned as discrete, static, and context free, and the latter as overlapped, dynamic, and context sensitive. The coarticulated nature of speech contributes to this theoretical puzzle, requiring perceptual constancy despite physical variations in the signal. Therefore, one "litmus test" for invariance is found in the place of articulation in stop consonant + vowel utterances, as these provide a complex example of the coarticulated nature of speech (Sussman *et al.*, 1991). The acoustic invariance theory makes two claims (Blumstein, 1986). First, there is acoustic invariance in the speech signal corresponding to phonetic features of natural language. It is hypothesised that the speech signal is highly structured in that it contains invariant acoustic patterns for phonetic features, and these patterns remain invariant across speakers, phonetic contexts, and languages. Second, the perceptual system is sensitive to these invariant properties. It is hypothesised that the perceptual system can use these patterns to provide the phonetic framework for natural language, and to process the sounds of speech in ongoing perception.

While numbers of studies have examined the invariance hypothesis for place of articulation in English stop consonants, there has been some disagreement about the particular form of the invariance. Some studies search for invariances in the formant transition region from stop to vowel, and others search for invariances in the region of the burst release.

3.3.1 Invariance: Formant Transition

The earliest contribution towards a theory of invariance was the concept of "hubs", which was first used by Potter *et al.* (1947). A "hub" is defined as the position of bar two (F_2 in Hz) when sound is produced in isolation. In their experiment using <*stop*><*vowel*> tokens it was observed that labial and alveolar stop consonants had their corresponding "hub" at the bottom and middle part of the spectrogram respectively. The velar "hubs" varied with the following vowel

context, such that the front, middle and back vowels had high, middle and low hubs respectively. The positions of labial and alveolar stop consonant hubs were postulated as invariant, but not so for velar stop consonants.

Later searches for invariance in the formant transition region were conducted by: synthesis and perception (Liberman *et al.*, 1954; Delattre *et al.*, 1955); vocal tract analogy (Stevens and House, 1956); analysis-by-synthesis (Stevens *et al.*, 1966), and analysis of natural speech (Kewley-Port, 1979, 1982). Sussman *et al.* (1991) analysed the "locus equations" as a quantitative index demonstrating relational invariant acoustic properties for place of articulation for the interval beginning with the F_2 transition and ending in the mid-vowel nucleus. The "locus equations" are formulated solely on relational differences in the second formant frequency across the time domain of the transition.

3.3.2 Invariance: Burst Release

Stevens and Blumstein (1978), and Blumstein and Stevens (1979) analysed burst spectra in an attempt to find invariant acoustic patterns for place of articulation. Three distinct spectral shapes were derived, *diffuse–falling*, or *flat*, for labials, *diffuse–rising* for alveolars, and a mid-frequency *compact* spectral peak for velars. These spectral templates categorised the initial stop place with approximately 85% accuracy. The spectral templates failed however, to adequately characterise place of articulation in languages other than English (Lahiri, 1980; Lahiri and Blumstein, 1981).

When Kewley-Port (1983) modified the fixed time window (26ms) of Blumstein and Stevens (1979), using a running spectral display advancing by 5ms steps to include the initial 40ms of the onset waveform, her three time-varying features displayed as visual patterns were categorically judged as /**b**/, /**d**/, /**g**/ with 88% accuracy.

Lahiri *et al.* (1984) used a new metric that quantified relative changes in distribution of energy from the burst release to the onset of voicing. They used these

metrics to categorise /**b**/, /**d**/ from three languages (English, French, Malayalam) of six speakers and obtained an average classification rate of 91% correct. This demonstrated relative differences of spectral energy levels as a place cue.

3.4 Sub-Phonetic Unit Modelling

The traditional units of modelling for automatic speech recognition have been word, syllable, phone, triphone, etc. This section briefly introduces some recent preliminary work on smaller units than those mentioned above.

Work on automatic phoneme recognition using distinctive features dates from Wiren and Stubbs (1956) who applied binary tree classification principles to the phoneme recognition problem. First they separated "voiced" sounds from "voiceless". Then they divided "voiceless" sounds into fricatives and voiceless plosives. They repeated this process until a single phoneme was isolated, with decisions being based on acoustic features present in the signal. For short vowels spoken by 21 speakers, they claimed 94% accuracy.

Bimbot *et al.* (1991) modelled mostly articulatory features using a Time Delay Neural Network (TDNN) with manually segmented continuous speech uttered by one French male speaker using the set of features summarised in Table 3.3. The results for feature extraction were in the range of 84.4% to 98.9% correct. Their conclusion was that the manner of articulation gave more accurate results than the place of articulation, because the acoustic correlates for manner of articulation are more directly rendered in the spectrum than are those for place of articulation.

Dalsgaard *et al.* (1991, 1992) used acoustic–phonetic features in a label alignment task in a multi-lingual environment (Danish, English, Italian). They used a Self-Organising Neural Network (SONN) to form the feature maps. The SONN transforms cepstral coefficients into a set of features, which is subsequently transformed into a set of principal components. These are used to model individual phonemes, which are used in a Viterbi search/level-building process to align an

Features
vowel/non-vowel
vocalic/non-vocalic
voiced/unvoiced
sonorant/non-sonorant
nasal/non-nasal
grave/acute
extreme/central
compact/diffuse
rounded/unrounded
flat (bemol)/non-flat (non-bemol)
delayed/non-delayed
discontinuous/continuous
fricative/non-fricative

Table 3.3: Features included in Bimbot *et al.* (1991).

Vowel Feature	*Consonant Feature*
sonorant	sonorant
front	plosive
central	fricative
back	nasal
high	liquid
middle	lateral
low	labial
round	dental
–	alveolar
–	velar

Table 3.4: Features included in Dalsgaard *et al.* (1991).

Feature
voiceless
noiseness
nasalness
frontness
centralness
backness
vowelness

Table 3.5: Features included in Elenius and Takács (1990).

independently given string of phonemes with the corresponding speech signal. The set of features they used is summarised in Table 3.4.

Zue *et al.* (1989), and Glass and Zue (1988) used a very different approach. They tried to find regularities in the acoustic signal by organising the so-called acoustic events which represent the changes of the acoustic signal in a multi-level structure. In this structure, information ranging from coarse to fine is represented in an organised fashion called a *dendrogram*. Two adjacent events form the onset and offset of a segment of a speech signal. Each segment is further classified as a phone. Phones can be transformed to phonemes at a later stage.

Elenius and Takács (1990) reported their results with recognition of phonemes in the continuous speech of a single speaker of Swedish and a single speaker of Hungarian. The recognition task was divided into two levels. At the first level, they used an MLP with one hidden layer to extract seven coarse features (Table 3.5) from 16 filter outputs of speech signal pre-processing. At the second level, seven sets of features extracted from the results of the previous level (associated with seven adjacent frames of signal together with the filter output of the central frame) are input to a second MLP, the output being the corresponding phoneme class. They reported that the coarse feature recognition rate was in the range of 80% to 95% accuracy and phoneme recognition accuracy was 64%.

Howard and Huckvale (1989) reported their result in a digit recognition task using a multi-speaker database. The acoustic signal was first pre-processed in a

Feature	*Description*
SIL	silence
FRIC	presence of frication
VOC	presence of voicing
NAS	presence of nasality
VFRIC	presence of voiced frication
S	presence of /s/ fricative
FTH	presence of /f/, /θ/ frication
?	glottal stop
K-REL	/k/ release
T-ASP	/t/ aspiration
EE-IH	/i/, /ɪ/ vowel
EH	/e/ vowel
UH	/ʌ/ vowel
ER	schwa vowel
AW	/ɔ/ vowel
UE	/ʊ/, /u/ vowel
R	/r/ glide

Table 3.6: Features included in Howard and Huckvale (1989).

bank of filters. Five frames of the pre-processed material were input to a feature extraction MLP, which classified the input into 17 features (Table 3.6). The first MLP's output formed input to the second MLP which used these features to recognise the digits. The feature extraction accuracy was in the range of 60.6% to 98.9%, leading to an overall digit recognition accuracy of 96.2%.

Later, Huckvale and Howard (1990) reported their results for phoneme recognition in a monosyllabic environment with a single speaker. At the first level of their system, a set of articulatory features plus some acoustic features were extracted using multiple MLPs (one for each feature), as summarised in Table 3.7. The phoneme recognition task was performed using HMMs with the extracted features as input. The recognition accuracy for 10 vowels was 44%, and for the 23 consonants was 17%, with 57% for 7 broad categories of consonants (voiced stop, unvoiced stop, voiced fricative, unvoiced fricative, nasal, liquid, and silence). The degradation of the results from their previous experiments may due to the much wider phonetic context.

Type	*Feature*
vowel	nucleus, close, open, front, back
broad manner categories for consonant	silence, unvoiced stop, voiced stop, voiced fricative, unvoiced fricative, nasal, liquid
consonant	envelope, nucleus, periodic source, aperiodic source, nasal

Table 3.7: Features included in Huckvale and Howard (1990).

Type	*Feature*
Manner of Articulation	closure, burst, vowel
Tongue body vertical	high, middle, low
Tongue body horizontal	front, middle, back
Place of articulation	labial, dental, front, velar, back velar, pharyngeal
Voicing	+, −

Table 3.8: Features included in Deng and Erler (1991).

With the aim of modelling coarticulation effects, Deng and Erler (1991) modelled a set of articulatory features (Table 3.8) using HMMs. They evaluated their approach in a speaker dependent CV syllable identification task, where C comprised six English stop consonants and V represented three cardinal vowels (/i/, /a/, /u/). The average identification rate achieved for the CV syllables was 96.23%

Stevens (1992) proposed a model of lexical access in which the acoustic properties are first extracted, these properties are associated with patterns of features and then these feature patterns are matched against lexical items (which are also represented in terms of patterns of features). The acoustic properties are extracted by identifying landmarks in the acoustic waves.

In summary, modelling of sub-phonetic units is at an exploratory stage. This modelling has been used in speech recognition and speech alignment. The sub-phonetic units have mainly concentrated on a somewhat arbitrary selection of phonetic features including articulatory features, acoustic description features

(describing the acoustic properties of the signal) and phonetic features (describing the type of sounds, e.g. glottal stop, frication, etc.), with emphasis placed on articulatory features. The features range from coarse to fine. This modelling approach has been used in different language environments (French, Swedish, Hungarian, and English), mostly in a speaker dependent mode. The techniques used for this modelling are SONN, TDNN, MLP and HMM. The different kinds of neural networks are used in a monolithic fashion where a single neural network is used to detect all the features in each case. The feature detection results are quite variable, ranging from 60% to 98%.

Some researchers used this sub-phonetic modelling for speech recognition with a monolithic approach. Some used the feature detection results as input to a monolithic speech recogniser; some used coarse feature detection results and spectral information of the incoming speech as input. The recognition results are more variable than the feature detection results.

3.5 Concluding Remarks

The overall impression arising from the above review of related research is that as yet we have a very limited understanding of the speech signal and that many questions remain to be answered.

The speech community is divided on the question of whether or not invariant cues exist in the acoustic signal. Although some researchers are intuitively convinced of the existence of invariant cues, they have not yet been able to establish a satisfactory quantitative description of invariances in the acoustic signal.

A further related point is that there are different types of phonetic features, whose nature is much more descriptive than quantitative. The search for acoustic correlates of those phonetic features continues. If we can find a suitable set of acoustic correlates of these features, we will have effectively provided evidence of invariances.

Finally, there is uncertainty regarding the nature of phonetic features, with different opinions as to whether features are binary or continuous in nature.

A few initial investigations of speech recognition using phonetic features have been reported, but this approach was taken only as an alternative to traditional pattern matching techniques. Questions that need further research effort are: (i) How can we use this approach to help find answers to some of the theoretical questions? (ii) How can we utilise the advantages of having a hierarchical system, in analysing errors made at intermediate levels and making necessary improvements?

A principal aim of this thesis is to investigate such questions further within the paradigm described in Chapter 2.

Chapter 4

Acoustic-Phonetic Transformation

This study investigates incorporation of distinctive features in a speech recognition system. The way in which the features are incorporated is by having an intermediate level where a transformation from the acoustic domain to a distinctive feature domain is achieved. Phoneme recognition is performed on the basis of the results of this transformation (see Figure 4.1). This chapter describes the techniques and phonetic features investigated in this thesis study. Chapter 6 describes a prototype recognition system in detail and Chapter 7 analyses transformation between acoustic and distinctive feature domain.

4.1 Feature Selection

From Figure 4.1, one can observe that the acoustic/feature transformation relates distinctive features to their acoustic properties, therefore, it is desirable to have a set of features parameters which describes acoustic properties in order to make this transformation conceptually feasible and to have a clear view of the correlation between the acoustic input and the corresponding set of features. One of the characteristics of Jakobson *et al.*'s (1961) features is that they are based on acoustic properties of the speech sounds (Section 3.2.3.2). Each feature is generally described in acoustic, articulatory and perceptual terms. This characteristic matches the requirements of the set of distinctive features needed for this study and we therefore adopt Jakobson *et al.*'s set.

In order to define distinctive feature values for Australian English, a more detailed review of the subset of Jakobson *et al.*'s distinctive features relevant to

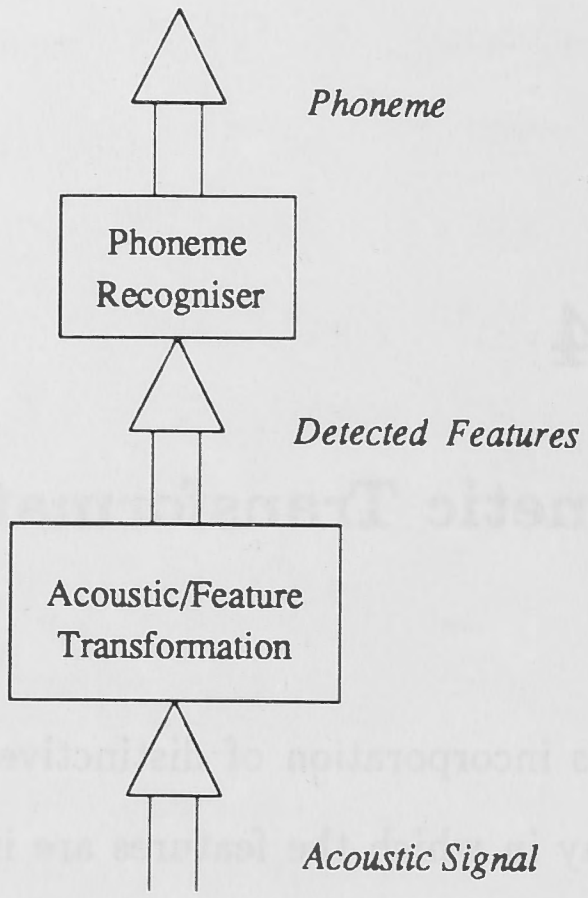


Figure 4.1: A global view of the system.

the speech material used in this study is presented in Section 4.3, emphasis being placed on the acoustic and articulatory properties of these features. Section 4.4 defines a specific subset of Australian English phonemes in terms of these features.

4.2 Technique Selection

In traditional knowledge–based speech recognition systems, explicit rules form the knowledge base. As indicated in Chapters 2 and 3, we do not fully understand the acoustic correlates of the individual distinctive features; how best to extract and to represent this knowledge is an open question. Section 2.5 described the consequent decision taken in this study to use an approach in which a knowledge–based approach is combined with pattern recognition techniques. These techniques are used as tools in the quest to gain a deeper understanding of speech signals, especially to gain insights into the way that phonetic information is represented in speech signals (and can thus be recovered). As we do not know what are the quantitative acoustic properties of the features, for the transformation from acoustic domain to the distinctive feature domain illustrated in Figure 4.1 to be feasible, we require a data driven technique which has a high

multi-dimensional mapping power, and can abstract from labelled examples in forming a mapping from acoustic parameters to distinctive features. This approach circumvents the problem of specifying speech knowledge as rules which set out the specific acoustic criteria for each of the distinctive features. The particular pattern recognition technique selected is Artificial Neural Network (ANN) of the Multi-Layer Perceptron (MLP) type, because their capabilities match the criteria mentioned above and they are known to be capable of performing complex non-linear pattern recognition tasks (Rumelhart *et al.*, 1986b). Section 4.5 reviews this type of ANN.

4.3 Review of Jakobson *et al.*'s Features

Jakobson *et al.* (1961) revised their work of 1952. They described their features in terms of sounds' acoustic properties, and where possible, their production. They divided features into inherent features and prosodic features. Inherent features include *acute/grave, compact/diffuse, voiced/unvoiced*, etc. This group of features does not change with the time, therefore they are static. Prosodic features are time dependent and therefore dynamic. The later are superposed upon the former and are lumped together with them into phonemes. Jakobson, Fant and Halle treated only inherent features in their *Preliminaries to speech analysis* (Jakobson *et al.*, 1961). Table 4.1 presents a summary of three types of inherent features in Jakobson *et al.*'s distinctive feature system.

This study considers only those features which are relevant to analysis of the speech material used herein. The relevant set of features happens to be the set of inherent features, and is reviewed in the remainder of this section.

- *Grave/Acute*

  Acoustic property: for *acute*, the upper formants of the spectrum predominates over the lower formants. For *grave*, the lower formants predominates.

Type	*Sub-category*	*Feature Type*	*Features*
Fundamental source feature		Vocalic	vocalic vs. non-vocalic
		Consonantal	consonantal vs. non-consonantal
Secondary consonantal source feature	Primary source	Envelope feature	continuant vs. interrupted
		Stridency feature	strident vs. mellow
	Supplementary source	Voicing feature	voiced vs. voiceless
Resonance feature	Basic resonator	Compactness feature	compact vs. diffuse
		Tonality feature	grave vs. acute flat vs. plain sharp vs. plain
		Tenseness feature	tense vs. lax
	Supplementary resonator	Nasalisation	nasal vs. oral

Table 4.1: Summary of Jakobson *et al.*'s inherent distinctive features.

The position of the second formant in relation to the other formants in the spectrum is the most salient property of this feature. When F_2 is closer to F_1, the sound is *grave*, when F_2 is closer to F_3, the sound is *acute*.

Articulatory Property: *grave* normally characterises the back vowels, labial and velar consonants. *Acute* is a property of the front vowels, dental and palatal consonants.

- *Compact/Diffuse*

  Acoustic property: for *compact*, a centrally (versus non-centrally for *diffuse*) located formant or set of formants predominates.

  The relation of the F_1 position to that of the higher formants is the characteristic aspect of this feature. If F_1 is closer to the higher formants, the sound is *compact*; otherwise it is *diffuse*.

  Articulatory Property: the ratio between the resonating cavity in front of the narrowest stricture and the one behind it determines this feature. The higher the ratio the more *compact* is the sound. *Compact* generally characterises open vowels, whereas *diffuse* relates to close vowels.

- *Flat/Plain*

  Acoustic Property: a downward shift (versus no downward shift for *plain*) of a set of formants or of all the formants.

  Articulatory Property: the *flat* feature is caused by reduction of the lip orifice, in other words lip rounding.

- *Lax/Tense*

  Acoustic property: for *lax* shorter (versus longer for *tense*) sound interval and lower (versus larger) energy.

  Articulatory Property: greater (versus smaller for *tense*) deformation of the vocal tract away from its rest position.

- *Vocalic/Non Vocalic*

  Acoustic Property: presence (for *vocalic*) versus absence (for *non-vocalic*) of a sharply defined formant structure.

  Articulatory Property: primary or only excitation at the glottis together with a free passage through the vocal tract for *vocalic*, additional noise source for *non-vocalic*.

- *Voiced/Unvoiced*

  Acoustic Property: presence (for *voiced*) versus absence (for *unvoiced*) of periodic low frequency excitation.

  Articulatory Property: periodic vibration of vocal cords (for *voiced*) versus lack of such vibration (for *unvoiced*).

The correlations of the distinctive features to the articulatory dimensions for vowels and for consonants are summarised by Hyman (1975) (see Table 3.1 for vowels and Table 3.2 for consonants). Figures 4.2 through 4.5 should help in visualising these correlations (Figure 4.5 relates features *tense/lax* to both acoustic and articulatory domain).

4.4 Feature Definition for Australian English

Section 3.2.3.6 argued that when defining feature values for phonemes, one should define them phonetically, i.e. according to the pronunciation of the sounds. Phonetically, Australian English varies to some degree from American and British English, although phonemically they are similar (Mitchell and Delbridge (1965)).

Most of the variations between Australian English and other forms of English occur in the vowels. The remainder of this chapter describes the derivation of the values of distinctive features for Australian English which corresponds to Jakobson *et al.*'s.

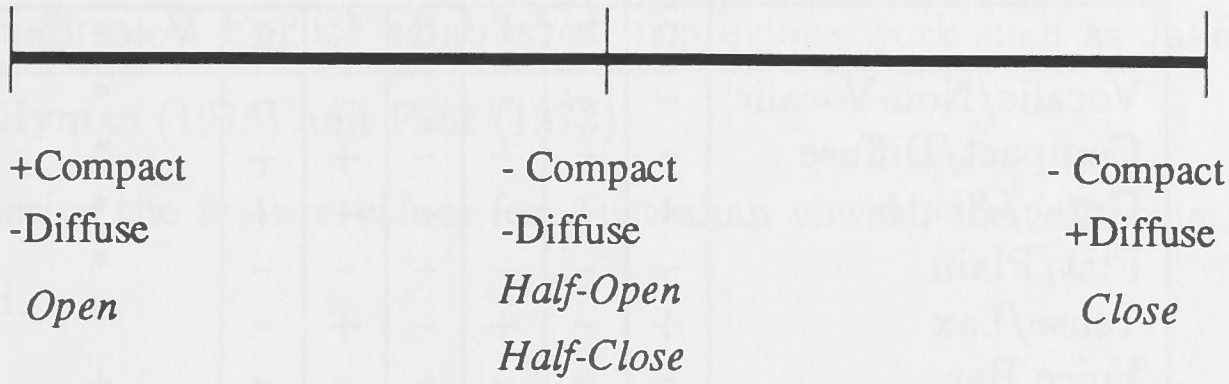


Figure 4.2: Correlation between *compact/diffuse* and vowels.

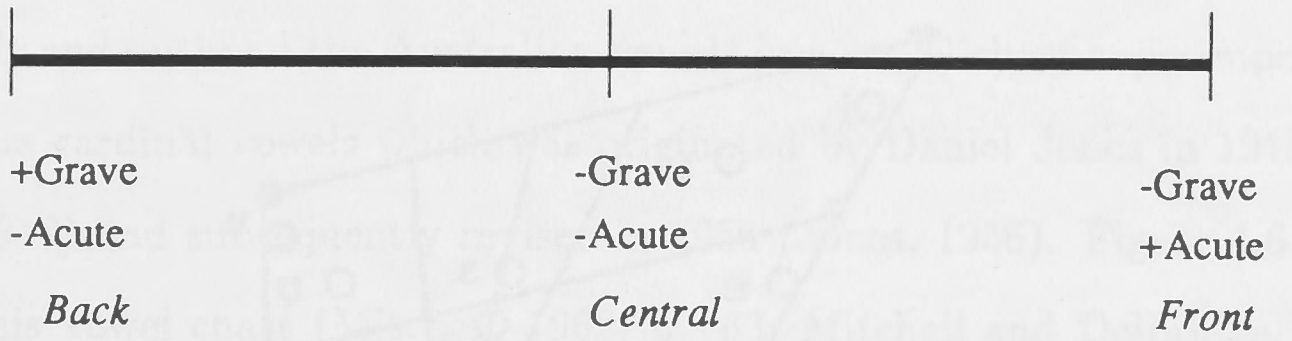


Figure 4.3: Correlation between *grave/acute* and vowels.

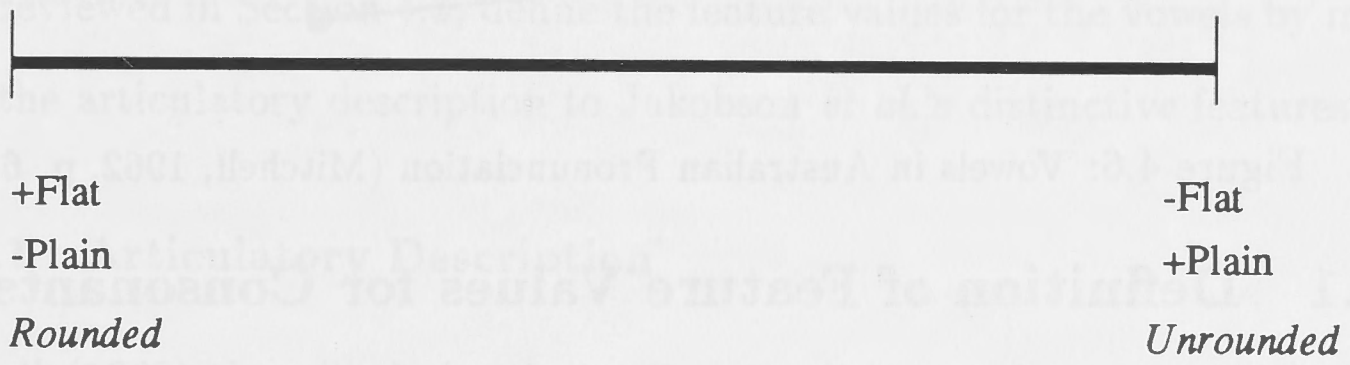


Figure 4.4: Correlation between *flat/plain* and vowels.

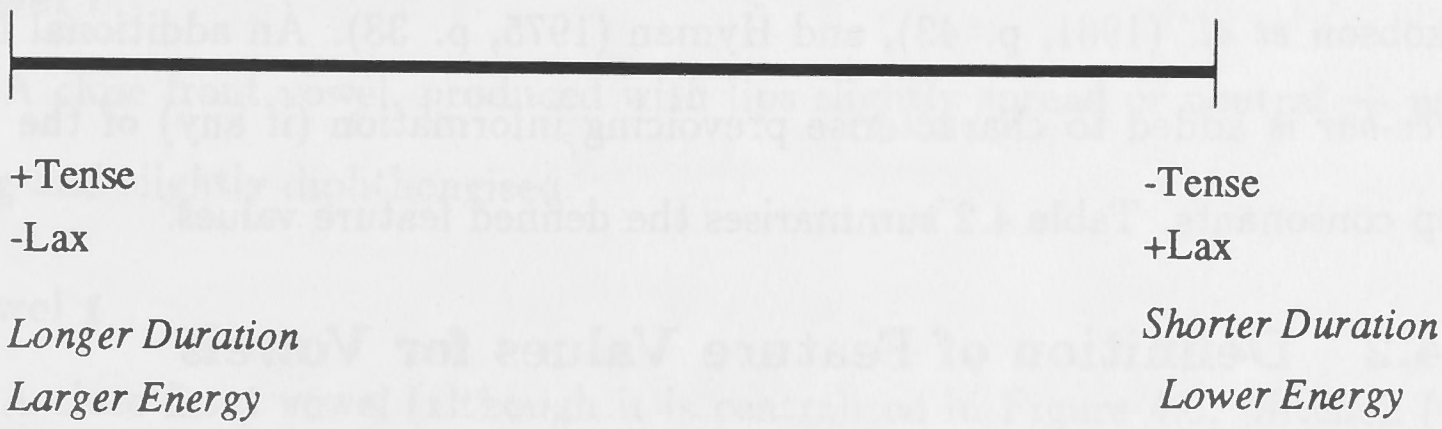


Figure 4.5: Correlation between *tense/lax* and vowels.

Features	*Stop Consonants*						
	p	**b**	**t**	**d**	**k**	**g**	Voice Bar
Vocalic/Non-Vocalic	−	−	−	−	−	−	*
Compact/Diffuse	−	−	−	−	+	+	*
Grave/Acute	+	+	−	−	+	+	*
Flat/Plain	−	−	−	−	−	−	*
Tense/Lax	+	−	+	−	+	−	*
Voice Bar	−	−	−	−	−	−	+

Table 4.2: *Feature values for stop consonants:* "+": feature is on; "−": feature is off; "*": feature does not apply to the given sound.

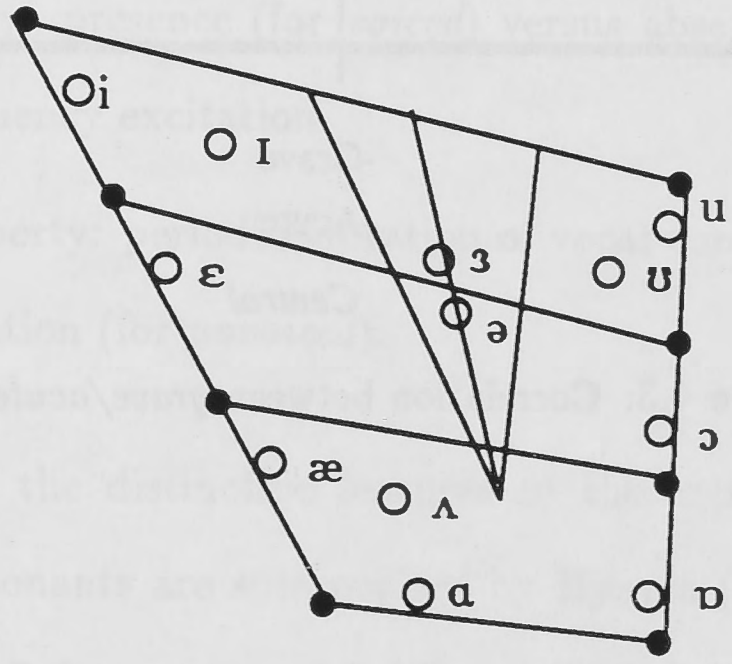


Figure 4.6: Vowels in Australian Pronunciation (Mitchell, 1962, p. 63).

4.4.1 Definition of Feature Values for Consonants

As the consonants in Australian English do not vary very much from other forms of English, e.g. American English, definition of the feature values for the stop consonants included in this thesis ([**b**, **p**, **d**, **t**, **g**, **k**]) is straightforward and follows Jakobson *et al.* (1961, p. 43), and Hyman (1975, p. 38). An additional feature *voice-bar* is added to characterise prevoicing information (if any) of the voiced stop consonants. Table 4.2 summarises the defined feature values.

4.4.2 Definition of Feature Values for Vowels

The Australian vowels included in this study comprise [i, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. These vowels differ in various degrees from those of other forms of English,

and in defining their feature values we did not directly borrow the definitions for other forms of English established in previous work such as Jakobson *et al.* (1961), Hyman (1975) and Fant (1973).

To derive the feature values for Australian vowels, the following mechanism was used.

- Define the Australian English vowels in their corresponding articulatory description, such as *close/open* (tongue high/tongue low) and *front/back.* Mitchell (1962) and Mitchell and Delbridge (1965) analysed Australian vowels and outlined the Australian vowels in a vowel chart superimposed with the cardinal vowels which was originated by Daniel Jones in 1918 (Jones, 1918) and subsequently revised in 1956 (Jones, 1956). Figure 4.6 presents this vowel chart (Mitchell, 1962, p. 63; Mitchell and Delbridge, 1965, p. 35).
- According to the description of features given by Jakobson *et al.* (1961) and reviewed in Section 4.3, define the feature values for the vowels by mapping the articulatory description to Jakobson *et al.*'s distinctive features.

4.4.2.1 Articulatory Description

Mitchell (1962) described the Australian English vowels in articulatory terms based on the speech of an educated speaker in Australia. These are summarised in the following.

Vowel i

A close front vowel, produced with lips slightly spread or neutral — usually long and slightly diphthongised.

Vowel ɪ

A close front vowel (although it is centralised in Figure 4.6, Mitchell (1962) still describes it as a front vowel), produced with the lips in neutral position — usually short.

Vowel ɛ

A half-close front vowel, produced with the lips in neutral position — usually short.

Vowel æ

A half-open front vowel, produced with the lips in neutral position — usually short.

Vowel ɑ

A front open vowel, produced with the lips in neutral position — usually long.

Vowel ɒ

A back open vowel, produced with the lips in neutral position — usually short.

Vowel ɔ

A half-open back vowel, produced with the lips in open-rounded position — usually long.

Vowel ʊ

A half-close back vowel, produced with the lips in close-rounded position — usually short.

Vowel u

A close back vowel, produced with the lips in close-rounded position — usually long, and slightly diphthongised. This description differs from Bernard's (1989) acoustic results; this point is referred to later when results are analysed.

Vowel ʌ

A front open vowel, produced with the lips in neutral position — usually short.

Vowel ɜ

A half-close central vowel, produced with the lips in neutral position — usually long.

The best descriptions available of these vowels in the *back/front* and *open/close* dimensions (Mitchell, 1962, p. 63) are presented in Figure 4.6.

Features	*Vowels*										
	i	ɪ	ɛ	æ	ɑ	ʌ	ɜ	u	ʊ	ɔ	ɒ
Vocalic/Non-Vocalic	+	+	+	+	+	+	+	+	+	+	+
Compact/Diffuse	–	–	-/-	-/-	+	+	-/-	–	-/-	-/-	+
Grave/Acute	–	–	–	–	–	–	-/-	+	+	+	+
Flat/Plain	–	–	–	–	–	–	–	+	+	+	–
Tense/Lax	+	–	–	–	+	–	+	+	–	+	–
Voice Bar	–	–	–	–	–	–	–	–	–	–	–

Table 4.3: *Feature values for vowels:* "+": feature is on; "–": feature is off.

4.4.2.2 Feature Derivation

Observing Figure 4.6, the horizontal dimension corresponds to *front/back*, the vertical dimension corresponds to *open/close*. More specifically, the upper left corner corresponds to *front/close*, the lower left corner corresponds to *front/open*, the upper right corner to *back/close*, the lower right corner corresponds to *back/open*. By mapping the articulatory dimensions to distinctive features, using the definition summarised in Section 4.3, the definition for vowels is derived, as summarised in the Table 4.3.

4.5 Artificial Neural Networks

Section 2.3.4 introduced Artificial Neural Networks (ANNs), and the related research on speech recognition using ANNs. This section reviews the type of ANN used for this study in more detail. ANNs have renewed interest in earlier work on perceptrons (Rosenblatt, 1959) which had lapsed, after Minsky and Papert (1969) proved that perceptrons could not make certain simple distinctions. This renewal of interest is a result of the development of new net topologies (e.g. Hopfield (1982)) and training algorithms (Rumelhart *et al.*, 1986b, Werbos, 1974; Le Cun, 1985; Parker, 1985). Lippmann (1987) also helped to publicise ANNs

with a lucid tutorial article. The intriguing analogy with biological information processing is captured in the name artificial neural network and this too helped attract interest.

There are several types of ANNs, such as the Multi-Layer Perceptron (MLP) (Rumelhart *et al.*, 1986b), Hopfield nets (Hopfield, 1982), Kohonen maps (Kohonen, 1984; Kohonen *et al.*, 1984), etc. MLPs have become widely used, and studied. Over the last few years, theoretical problems related to MLPs have been investigated by many researchers, the results being summarised by Hush and Horne (1993). As MLPs are the basis for experimental investigation described in subsequent chapters (for reasons described in Section 4.2), they are briefly reviewed here.

4.5.1 The Multi-Layer Perceptron

The MLP is one type of ANN, sometimes referred to as a feed-forward neural network. An MLP is constructed from an input layer, one or more hidden layers, and an output layer, with each layer having one or more nodes. The nodes of each layer can be fully or partially connected to the nodes of the subsequent layer, progressing from the input to the output layer. Figure 4.7 is a schematic view of a 13-4-2 MLP, which consists of 13 input nodes, one hidden layer of four nodes and two output nodes. Figure 4.8 portrays a single node, where

$$Y = f\left(\sum_{i=0}^{N-1} w_i x_i - \theta\right)$$

$f(\cdot)$ is called the activation function. Figure 4.9 shows two examples of these functions, where Figure 4.9a shows a hard-limiting nonlinearity:

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise} \end{cases}$$

and Figure 4.9b shows the sigmoid function:

$$f(x) = \left(1 + e^{-\beta x}\right)^{-1}$$

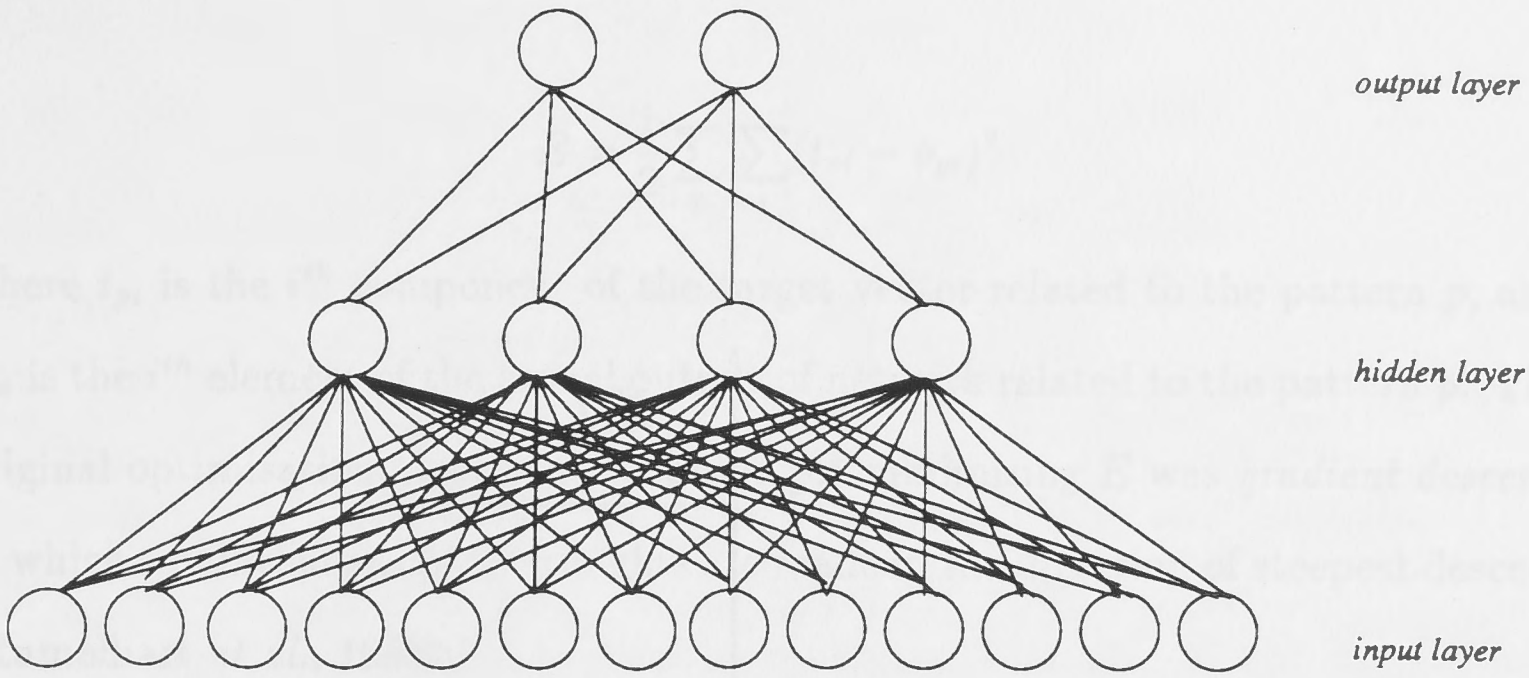


Figure 4.7: Schematic view of a Multi-Layer Perceptron.

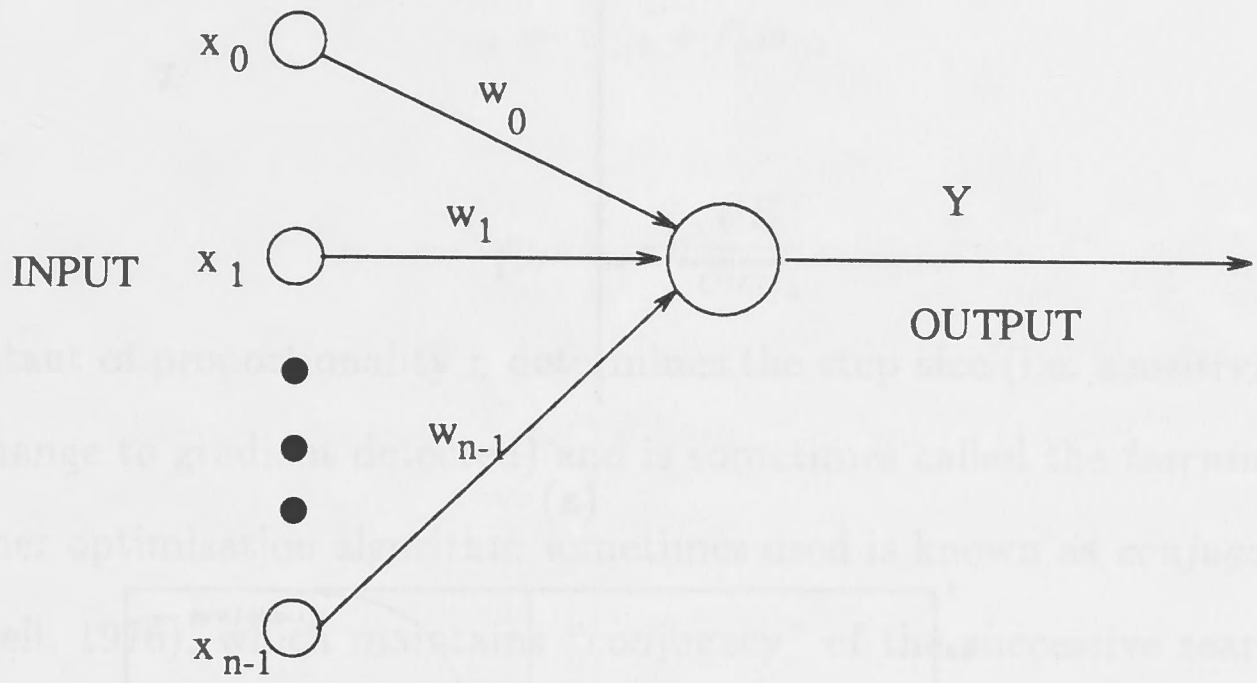


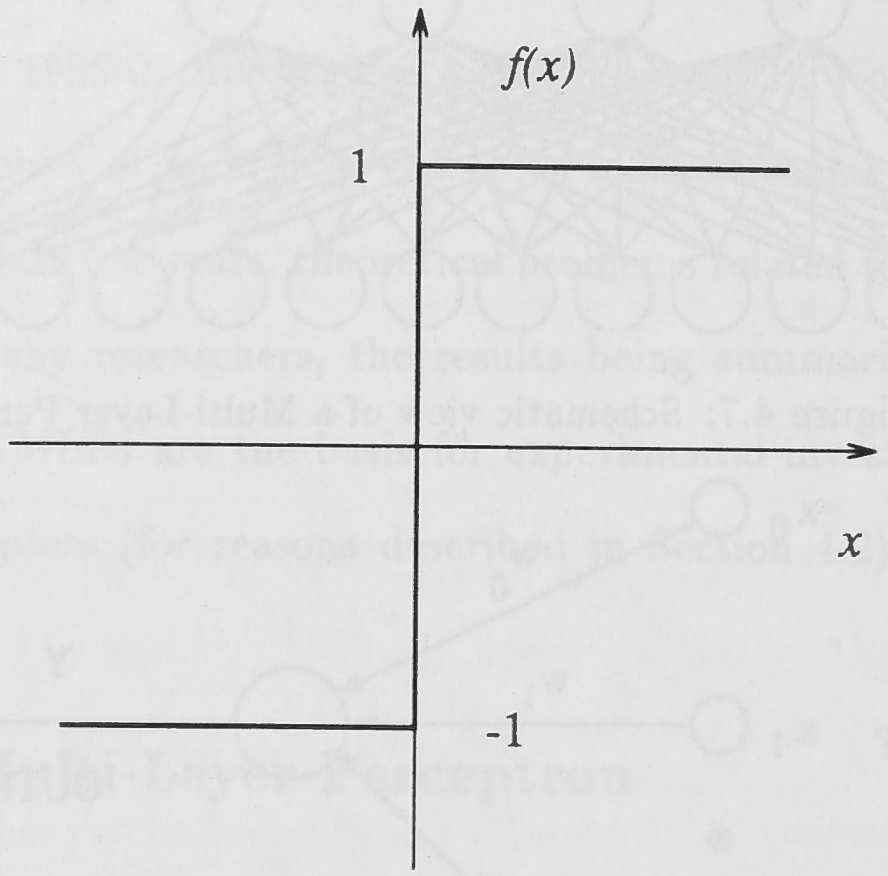
Figure 4.8: Schematic view of a single node.

where $\beta = 1.0$ in the case illustrated.

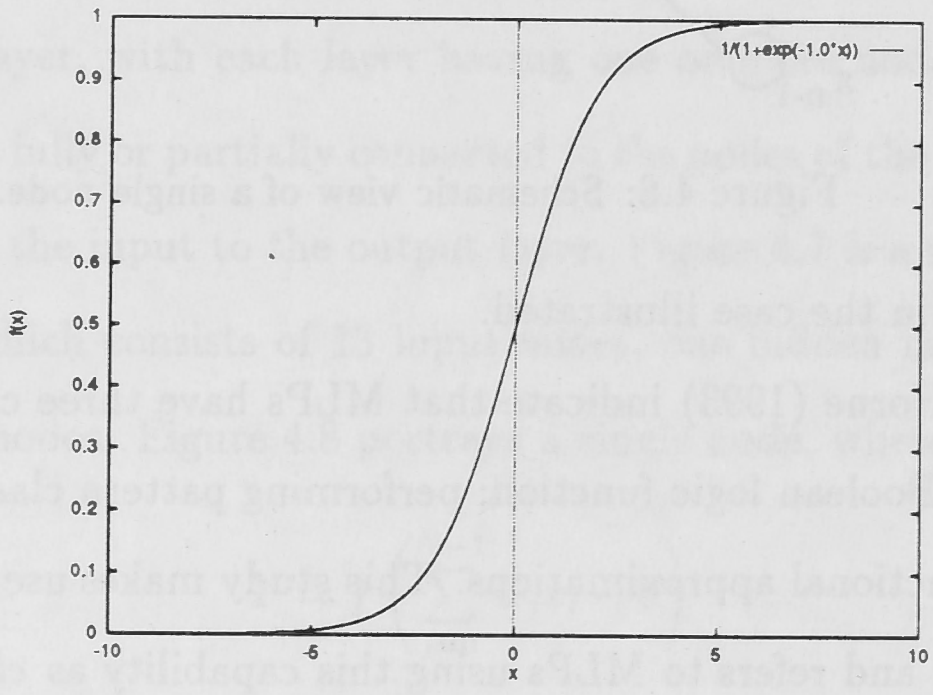
Hush and Horne (1993) indicate that MLPs have three capabilities, namely: implementing Boolean logic function; performing pattern classifications; and implementing functional approximations. This study makes use of the second capability of MLPs and refers to MLPs using this capability as *classifiers.*

4.5.2 MLP training

Rumelhart *et al.* (1986b) originally introduced *back-propagation.* The training procedure involves presentation of a set of pairs of input and output patterns. The system first uses the input vector to produce its own output vector and then compares this with the desired output or *target* vector. Rumelhart *et al.* propose that the following error measure function be optimised:



(a)



(b)

Figure 4.9: Activation functions: (a) Hard-limiting; (b) Sigmoid.

$$E = \frac{1}{2}\sum_{p}\sum_{i}(t_{pi} - o_{pi})^2$$

where t_{pi} is the i^{th} component of the target vector related to the pattern p, and o_{pi} is the i^{th} element of the actual output of network related to the pattern p. The original optimisation algorithm proposed for minimising E was *gradient descent*, in which small changes in the weights are made in the direction of steepest descent (Rumelhart *et al.*, 1986b):

thus

$$w_{ijk} \leftarrow w_{ijk} + \triangle w_{ijk}$$

where

$$\triangle w_{ijk} = -\eta \frac{\partial E}{\partial w_{ijk}}$$

The constant of proportionality η determines the step size (i.e. sensitivity of the weight change to gradient detected) and is sometimes called the *learning rate.*

Another optimisation algorithm sometimes used is known as *conjugate gradient* (Powell, 1976), which maintains "conjugacy" of the successive search directions when a quadratic function is optimised. This algorithm does not require any explicit second derivatives, simplifying the computations, and also uses less storage than the gradient descent algorithm (Barnard and Cole, 1989, Powell, 1976). Barnard and Cole (1989) further indicated that the *conjugate gradient* algorithm is usually able to locate the minimum of a multivariate function much faster than the *gradient descent* procedure.

4.5.3 Viewpoint: Mapping Functions

One way to view feed-forward networks is to regard them as mapping functions from an input space to an output space (Nilsson, 1965). For example, an MLP of one hidden layer with l input nodes, m hidden nodes, and n output nodes, has an l-dimensional input space I and an n-dimensional output space, O. An MLP is thought of as a mapping of a point in the space I to another space O:

$$ff : I \mapsto O$$

Since this MLP has one hidden layer that can be described as an m-dimensional space H, the mapping function ff is a composition of two functions:

$$ff = h \circ g$$

where

$$g : I \mapsto H, h : H \mapsto O.$$

For a given MLP, this multi-dimensional mapping function is not unique. The topology of the network determines the range of possible mapping functions it can represent. Training such a network implies the selection of a particular mapping. Training is accomplished by providing an appropriate set of labelled training patterns (which can be viewed as a *specification* of the desired mapping function). Cybenko (1989) has shown theoretically that a two-layer network (with one hidden layer) can form an arbitrarily close approximation to any continuous non-linear mapping. In practical situations, Chester (1990) demonstrated that for some problems, a small 3-layer network can be used where a 2-layer network would require an infinite number of nodes. Hajnal *et al.* (1987) showed that there are problems which require an exponential number of nodes in a 2-layer network that can be implemented with a polynomial number of nodes in a 3-layer network.

4.5.4 Viewpoint: Decision Boundaries

In contrast to the above view of MLPs implementing a multi-dimensional mapping function, this section reviews MLPs from the pattern classification standpoint.

When an MLP is used as a classifier, it maps an input space I of dimension m onto an output space O of dimension n, where m represents the number of elements in each input pattern vector to be classified and n represents the number

of classes to which the input patterns are assigned. An MLP performs this classification task by forming a hyperplane decision boundary with each hidden node. The hidden nodes divide the input space by such hyperplanes, forming decision regions. Lippmann (1987) demonstrated that a two-layer (i.e. one hidden layer) MLP can implement arbitrary convex boundaries, where each hidden node with a hard-limiting nonlinearity forms a hyperplane. Later, Makhoul, El-Jaroudi and Schwartz (1989) showed that two-layer networks can form an arbitrarily close approximation to any non-linear decision boundary through use of sigmoid non-linearities in the nodes. Cybenko (1989) reports similar findings.

4.5.5 Viewpoint: *A Posteriori* Probability Emitter

Richard and Lippmann (1991) shows that when a mean-squared-error criterion is used for the training an MLP, and 0's and 1's are used as target outputs, the Bayesian probability:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

is estimated. In this equation, $P(X|C_i)$ is the likelihood or conditional probability of producing the input X if the class is C_i, $P(C_i)$ is the *a priori* probability of class C_i, and $P(X)$ is the unconditional probability of the input (Duda and Hart, 1973; White, 1989). Richard and Lippmann (1991) demonstrated that many neural network classifiers provide outputs which estimate Bayesian *a posteriori* probabilities, using the following:

- mean-squared-error cost function;
- one of M outputs is unity, with the remainder zero as output targets for training.

More specifically, when we train an MLP with the back-propagation algorithm, the MLP can learn the best mean-squared-error approximation to the *a posteriori* probability (Hush and Horne, 1993).

4.5.6 MLP Architecture, Training Data Size and Generalisation

Lippmann (1987) demonstrated that each node in the hidden layer(s) forms a half-plane region. Convex regions are formed from intersection of such half-plane regions. More complex decision boundaries are formed by using more nodes. A more complicated architecture such as a fully connected MLP has more parameters (i.e. weights) to estimate. The estimation of these parameters requires larger training samples.

For a given problem, how complex must the architecture be to solve the problem? The answer is still unknown (Lippmann, 1987; Hush and Horne, 1993). If the architecture of an MLP is too simple, it may not be able to solve the problem; if it is too complex, it needs a large training data set to estimate the parameters adequately. If the training data set is too small for the complexity of architecture, the MLP may be able to implement *numerous solutions* that are consistent with the training data, but most of these are likely to be poor approximation to the actual problem (Hush and Horne, 1993). Hush and Horne (1993) suggest that with little knowledge of the problem, one must determine the network size by trial and error. They further suggest a strategy of starting with the smallest possible network and gradually increasing the size until the improvement in performance begins to level off. In this approach, each size of network is trained independently. Our strategy of handling this problem happened to be the same as Hush and Horne proposed.

Another related issue is *generalisation* which is a measure of how well the network performs on the actual problem once training is complete (Hush and Horne, 1993). Hush and Horne indicate that generalisation is influenced by the following parameters: the number of data samples for training (and how well they represent the problem at hand), the complexity of the underlying problem, and the network size. They suggest that for a network of the proper size, a

larger number of data samples will do a better job at representing the underlying problem. This should allow the MLP to learn a better solution to the problem.

Another practical issue in relation to training algorithms for MLPs is that they do not guarantee to find the global minimum in the error function being optimised. To avoid using local minima, Lippmann (1987) suggests to make many training runs starting with different sets of small random weights for the links, and selecting the best overall set of weights.

In summary, although there has been great interest in investigation of ANN theories, as well as applications, important problems remain to be solved, such as network architecture selection and local minimum. This thesis study utilises ANNs as a data driven technique matching our technique selection criteria described in Section 4.2 and does not aim to solve these outstanding problems. They are mentioned here to acknowledge their existence and to justify some of the ways in which they are used in this study. We follow the suggestions made by other researchers in the field described herein to overcome the related practical problems. In later chapters, when we report the results, we include the architecture selections to provide them as reference data to later studies.

Chapter 5

Speech Material and Acoustic Pre-processing

5.1 Speech Material

The speech material in this study concentrates on stop consonants and the nominally monophthongal vowels in Australian English. Coarticulation effects are evident in the <*stop*><*vowel*> syllables (Sussman *et al.*, 1991), making such data suitable for studying the robustness of distinctive feature extraction in the context of these coarticulatory variations.

Section 3.3.2 reviewed the research effort in searching for invariant cues in the burst release. In a typical example of such researchs, Stevens and Blumstein (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979) analysed burst spectra in order to find invariant acoustic patterns for place of articulation in <*stop*><*vowel*> context. They derived three distinct spectral shapes: *diffuse-falling* for labials, *diffuse-rising* for alveolar, and *compact* for velars. These spectral templates categorised the initial stop place with approximately 85% accuracy. The spectral shapes correlate with Jakobson *et al.*'s distinctive features as being *diffuse* and *grave*, *diffuse* and *acute*, and *compact* respectively. One of the reasons for selection of <*stop*><*vowel*> speech material is to compare the results of this study with those of Stevens and Blumstein.

The speech material includes <*stop*><*vowel*>[*d*] words where

<*stop*> = voiced and voiceless stop consonants = /**p**, **t**, **k**, **b**, **d**, **g**/;

< *vwl* > = Australian monophthongs = /**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**/.

This material was collected, digitised and segmented into words by our former colleague Frantz Clermont. The material was recorded from four native Australian English speakers in citation form (one session for each speaker with five repetitions). The words were presented to the speaker in random order of consonantal context for each vowel. Each word was presented visually at 3 sec intervals and the speaker was encouraged to take a fresh breath after each utterance.

The recording was performed in an acoustically-treated recording booth. Each word was visually presented to the speaker through a small window between the booth and the experimenter's area. The recording apparatus (Reel-to-Reel Tape Recorder, model REVOX A77) which had been previously calibrated was located in the experimenter's area.

The analogue inputs were amplified, low-pass filtered to 4.5 kHz using a Rockland amplifier/filter (Model 1042F), then sampled at a rate of 10,000 12-bit samples per second. Clermont and Millar (1986) summarised the recording procedures briefly. Clermont (1991) described more fully the details of the recording procedure, equipment used, and speaker selection. Clermont's (1991) description related to $[h]$<*vowel*>$[d]$ words, but the same conditions applied for the <*stop*><*vowel*>$[d]$ words which are of interest for this thesis.

5.2 Acoustic Signal Segmentation and Labelling

The departure point of this study is analysis of the burst of the stop consonants, and the pseudo steady-state vowels. Therefore, the speech material needed underwent a segmentation and labelling process.

The data was hand segmented and labelled, with each word being segmented into a voice bar interval (if present), a burst-onset interval, a transition interval from the stop consonant to the vowel, a pseudo steady-state vowel interval and a transition interval from the vowel to the final $[d]$.

As reported in Ran and Millar (1992), the segmentation was based on visual inspection of the spectrogram of the <*stop*><*vowel*>[*d*]'s, generated in an ESPS XWAVES version 1.1 environment. The criteria used in this segmentation (Ran and Millar, 1992) were:

- (For *[Voiced stop][vowel][d]* words) The starting point of the prevoicing; the indication of this point is the presence of a voicing bar;

- End point of the prevoicing and starting point of the burst; the indication of this point is the ending of a voicing bar and/or the starting of broad-band noise;

- End point of the burst and starting point of the aspiration (if present) or the transition from the stop to vowel; the indication of this point is the ending of the broad-band noise of high energy and/or the appearance of a relatively clear formant structure;

- End point of the transitional portion and starting point of the pseudo steady-state portion of the vowel; the indication of this point is the starting point of the relatively static formant structure;

- End point of the pseudo-static portion of the vowel and the starting point of the transitional portion from vowel to *[d]*; the indication of this point is the ending of the static formant structure and the starting point of the dynamic formant structure;

- Starting point of the *[d]* closure; the indication of this point is the absence of any formant energy described in the previous point and presence of a voicing bar for the *[d]*.

Feature	*Description*
Acute	Upper side of the spectrum predominates
Compact	Relative predominance of one centrally located formant region (or formant)
Diffuse	One or more non-central formants or formant regions predominate
Flat	A set of formants (or even of all formants) in the spectrum shifts downward
Grave	Lower side of spectrum predominates
Lax	Shorter sound interval and lower energy compared with Tense
NonVocalic	Having more than one periodic source whose onset is abrupt
Plain	No shift of formants
Tense	Longer sound interval and a larger energy compared with Lax
Vocalic	Having a single periodic source whose onset is not abrupt
VoiceBar	Presence of low frequency spectral energy
Voiced	Superposition of a harmonic sound source upon the noise source
Voiceless	Having noise source only

Table 5.1: Brief description of Jakobson *et al.*'s distinctive features.

5.3 Acoustic Analysis

5.3.1 Parameter Type Decision

Through analysis of the acoustic properties of the features described in Section 4.3, as summarised in Table 5.1, we can conclude that all features of interest to this study relate to the distribution of energy in the frequency domain, except for the pair of features *tense/lax* which are more duration and energy related in the time domain. Therefore, a form of parameterisation of the acoustic signal which represents the energy distribution in the frequency domain was adopted. There are few alternatives which fulfil these requirements, such as speech production

modelling based methods (e.g. Linear Predictive method), signal processing based methods (e.g. FFT based methods). Because LPC gives smoothed spectral representation and the cepstral coefficients perform better than the straight coefficients,which is commonly recognised fact, Linear-Predictive Cepstral Coefficients (LPCC) were selected.

5.3.2 Acoustic Parameterisation

The procedure for deriving the Linear-Predictive Cepstral Coefficients (LPCCs) is implemented in the ESPS environment. Frames of 12.8 ms were first passed through a Hamming window, with adjacent frames having 6.4 ms overlap. A set of reflection coefficients was calculated using the autocorrelation method. This set of coefficients was then transformed to autoregressive filter coefficients and finally, into cepstral coefficients. 13 LPCCs were derived for each frame. The analysis was performed on the prevoicing interval, the burst-onset interval and the pseudo-static vowel interval.

Chapter 6

Speech Recognition Based on Distinctive Features

6.1 Design of Experiments

The aims of the experiments described in this chapter are to investigate: (i) the feasibility of distinctive feature extraction in a speaker independent mode; (ii) how well the extracted features can be applied to speech recognition.

The system investigated here divides the ASR problem into two main levels. The first level detects distinctive features (described in Section 4.3) from the input data, in a parallel fashion. The second level combines the results of feature detection for phoneme recognition. The first level is vitally important because the second level is totally dependent on it. In the current chapter, Section 6.2 presents the results of an investigation into two schemes for feature detection, Section 6.3 describes the overall system design, Sections 6.4 and 6.5 describe the feature detection process in detail, and Section 6.6 describes phoneme recognition based on the detected distinctive features.

In order to evaluate the effectiveness of the feature detection level, two experiments were performed in which speech data are input to the system directly for recognition as described in Section 6.7. Section 6.8 compares the results of different approaches and summarises and discusses the results of the earlier sections.

6.1.1 Architecture Selection

MLPs are the primary tools used in the experiments described in this chapter. The architecture for each MLP used has to be defined. In selecting MLP architectures we take the following points into account (see Section 4.5.6):

- Small networks use simple decision boundaries formed in accordance with general properties of the training samples.

- Earlier research suggests that for good problem generalisation, the training data should be representative. If we cannot control the coverage of the training data, larger training data sets are better than small ones.

From Section 4.5, it is clear that theoretical questions of how to select the network size and training data for a given problem remain unsolved. As Section 4.5.6 indicated, this study does not intend to solve ANN problems, but takes a policy which concurred with the suggestion made by Hush and Horne (1993), namely to select the network size as small as possible in order to keep the ratio of the training data size to the free parameters of the network as large as possible. In this way, we hope to achieve good generalisation, where the resulting trained networks make decisions according to general properties present in the training data, rather than the detailed properties associated with individual training data samples. Architecture selection proceeds as follows:

1. Start with a network which has a small number of hidden units;

2. Train the network;

3. Evaluate the network with test data; if the performance is not satisfactory (accuracy $< 98\%$), increase the number of hidden units, and go back to 2;

This process is repeated until the performance does not improve noticeably by including additional hidden units. If the final performance is unsatisfactory,

(e.g. the accuracy is very much lower than the accuracy for other arrangements of training-testing speakers), then another hidden layer is added, and this process is repeated again.

The following sections report results only for those architectures which give best performance. The structure of the ANNs is reported for reference only, usually without comment.

6.1.2 MLP Training and Testing

The particular type of MLP used in this study is a classifier which has separate training and operational (testing) modes. During training, each input vector has a corresponding target output vector where each element of the output target vector represents one class. Each input vector can have only one class "on" (or "selected" by having a value of '1' for the corresponding element in the target vector) and the remaining classes "off" ("unselected" by having a value of '0' for all other elements in the output vector). Pairs of input and desired output vectors are presented to the MLP one by one, until a complete pass is made through all the training patterns. The *back-propagation* training algorithm is used to update weights from their initial values. This process is repeated until a minimum overall error is found which cannot be reduced further (or only very slowly). During this process, the MLP "learns" to represent the desired relationship between the input and output by its internal parameters. It learns from the data by extracting the information that best represents the input to output mapping and gradually accumulates this information in the network.

We follow Lippmann's (1987) suggestion of having many runs with different settings of initial small random values of weights (see Section 4.5.6). We had 100 runs with different initial settings and found that 100 trials was generally large enough to enable us to find one or more starting points which led to good solutions.

During operation, an input vector is presented to the trained network and the network calculates an appropriate output based on activation scores from each

output node. The input is then assigned to the class whose corresponding output node has the highest activation score.

6.2 Two Schemes of Feature Detection

As Section 6.1 indicated, because recognition is based on the feature detection results (see Figure 4.1), feature detection is vitally important. This section investigates two different schemes for feature detection. The central question is: As Jakobson *et al.* (1961) classified vowels and consonants within a single distinctive feature framework — are the corresponding acoustic properties of the distinctive features the same for vowels and for stop consonants?

The two schemes are named scheme A and scheme B. In scheme A, data from the burst-onset interval of the stop consonants, and data from the pseudo steady-state vowel interval and pre-voicing interval are aggregated in the same input data domain during training, and the acoustic properties of the features are extracted from this common data domain. In scheme B, acoustic properties of the features are extracted from pseudo steady-state vowels and the burst-onset interval of stop consonants separately. Ran and Millar (1993a) reported the results of an investigation of these two schemes for feature detection.

6.2.1 Scheme A

The acoustic properties of the features are extracted from the burst-onset interval of the stop consonants, the pseudo steady-state vowels, and the pre-voicing material together. One MLP of type classifier is used to detect presence/absence of each feature, as illustrated in Figure 6.1. During the training process, speech material consisting of three repetitions of burst-onset interval of stop consonants, one repetition from the pseudo steady-state vowel and one repetition of pre-voicing material from each of the three training speakers was used. Note that two more repetitions of the burst-onset interval were included to balance the total number of samples from each type of material (i.e. burst-onset interval and steady-state

vowel) for the MLP to learn from representative samples of each type of material. The number of samples of burst-onset interval material of each repetition is much smaller than the number of samples of pseudo steady-state vowel or pre-voicing material, because the burst-onset interval is much shorter than the other two types of signal. The input vectors were formed from 13 Linear Predictive Cepstral Coefficients (LPCCs) for each speech frame. There are three output nodes (i.e. classes) for some feature detectors where one node corresponds to the *on* class, one node corresponds to the *off* class, and the third node corresponds to the *irrelevant* class (meaning the feature does not apply to the corresponding input vector). The other feature detectors have two output nodes corresponding to *on* and *off* classes. The target vectors are derived from Table 4.2 for the burst-onset interval input material and pre-voicing material and from Table 4.3 for the pseudo steady-state vowel material.

The testing (operational) mode uses speech data from burst-onset intervals, pseudo steady-state vowel intervals, and pre-voicing intervals from a randomly selected speaker who was not included in the training phase.

Tables 6.1 through 6.3 report the results for feature detection obtained in the testing phase. Table 6.1 reports the best architecture of the MLPs found for each of the feature detectors based on the test performance. The notation for the architectures is *<number of input nodes>-<number of hidden nodes>[- <number of hidden nodes in the second hidden layer>]-<number of output nodes>*. For example, 13-14-3 means that there are 13 input nodes, 14 hidden nodes and 3 output nodes. Ran and Millar (1992) reported the results of these experiments which were conducted on a Sun 4.

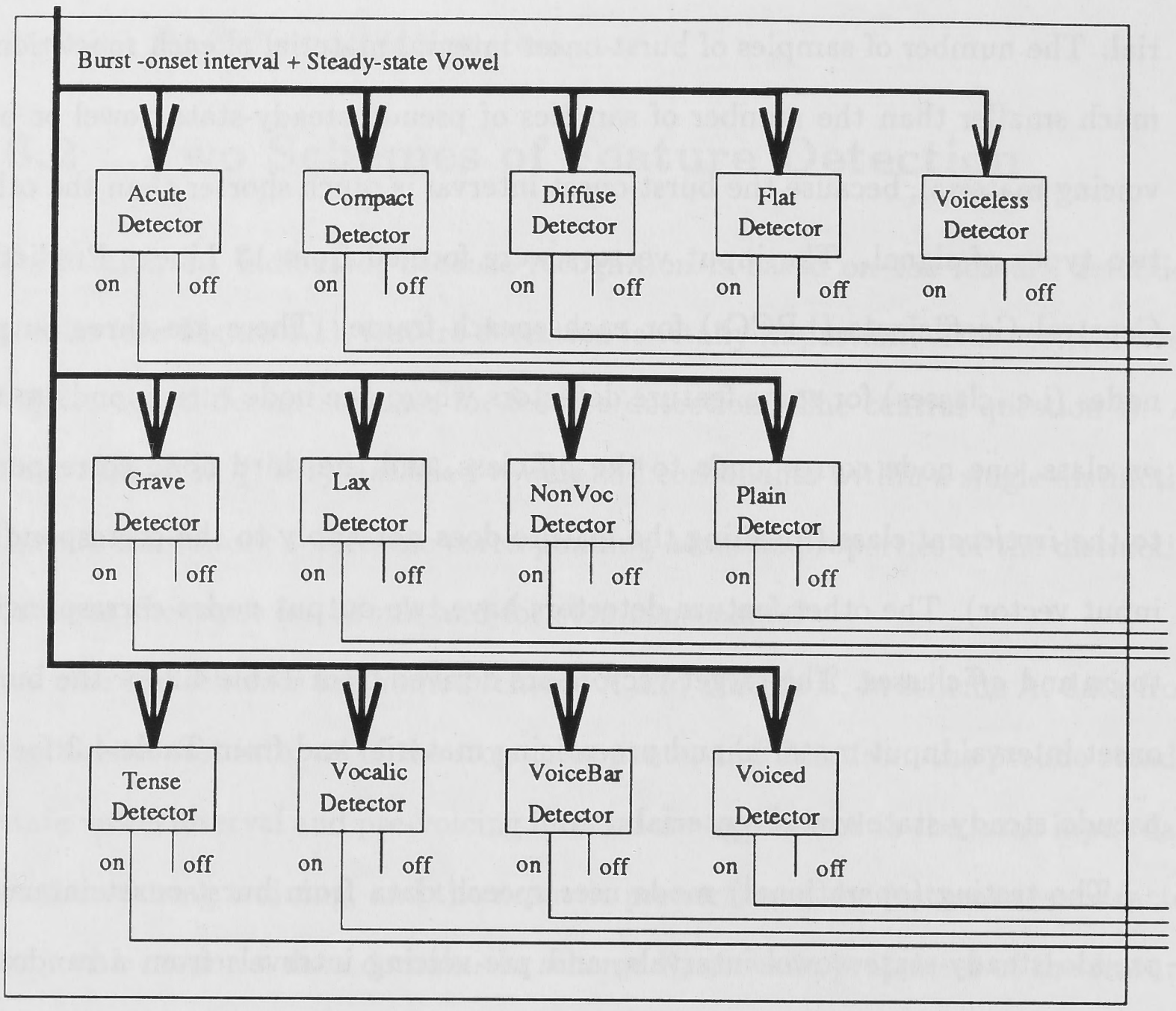


Figure 6.1: Structure of scheme A.

Feature	*Architecture*
acute	13-14-3
compact	13-12-3
diffuse	13-12-3
flat	13-12-3
grave	13-10-3
lax	13-12-3
non-vocalic	13-6-4-3
plain	13-8-4-3
tense	13-12-3
vocalic	13-6-4-3
voice bar	13-8-3

Table 6.1: Best MLP architectures for scheme A.

Feature	*Rate %*
acute	83.8
compact	90.1
diffuse	73.8
flat	89.6
grave	84.1
lax	80.8
non-vocalic	98.8
plain	88.5
tense	79.4
vocalic	94.9
voice bar	100.0

Table 6.2: *Scheme A:* Feature detection rate for pseudo steady-state vowels.

Feature	*Rate %*
acute	82.1
compact	90.4
diffuse	79.3
flat	97.8
grave	77.2
lax	54.1
non-vocalic	72.8
plain	98.1
tense	54.9
vocalic	76.9
voice bar	98.3

Table 6.3: *Scheme A:* Feature detection rate for burst-onset interval of stop consonants.

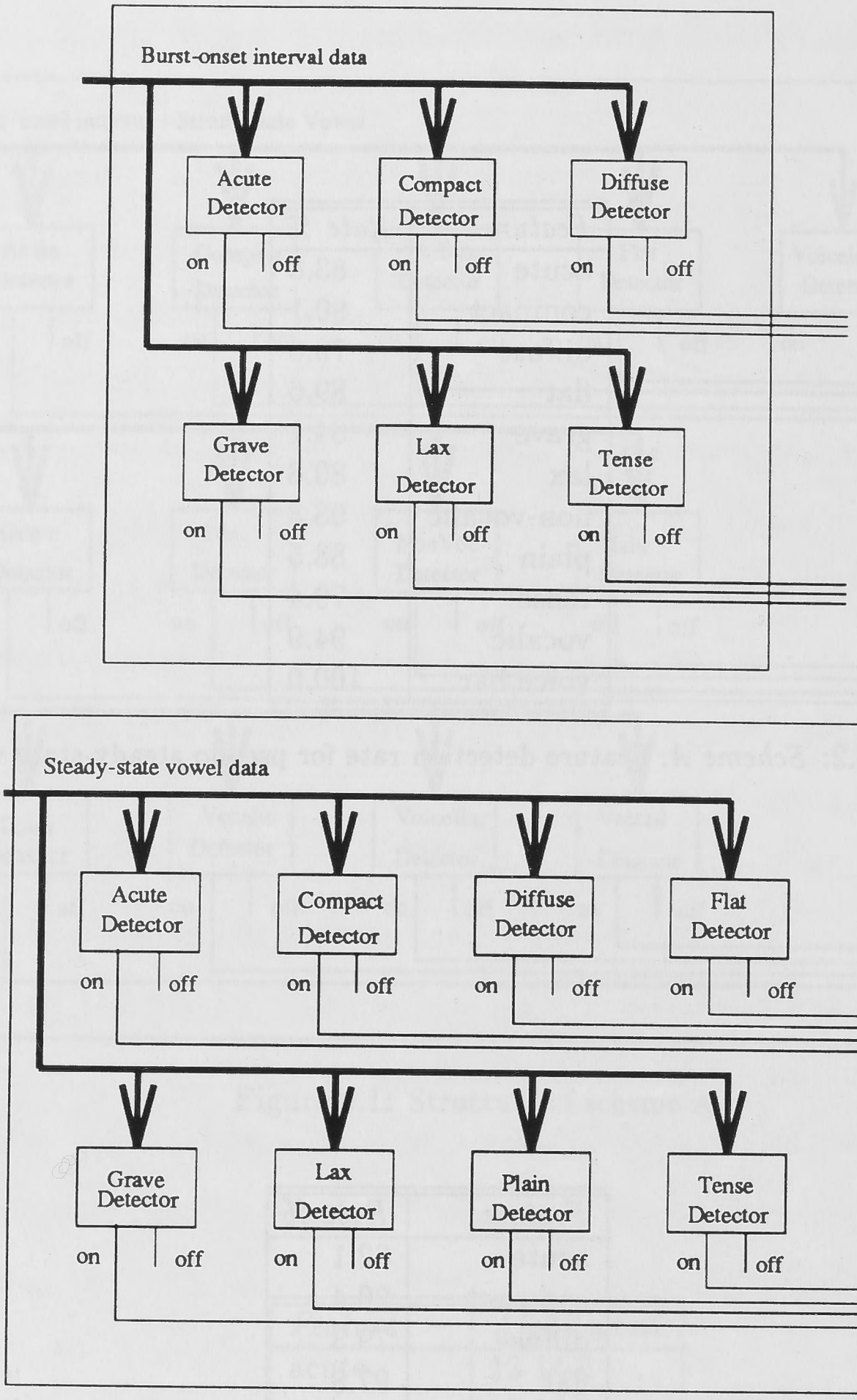


Figure 6.2: *Structure of scheme B:* (a) Feature detection for burst-onset intervals; (b) Feature detection for pseudo steady-state vowels; (c) Common feature detection.

6.2.2 Scheme B

Figure 6.2 illustrates the design of scheme B in which feature information for stop consonants and vowels is extracted separately (except for those common features). Three repetitions of burst-onset intervals of stop consonants from each of the three training speakers were used as training material for feature detection for the stop consonants. One repetition of pseudo steady-state vowels from each of the three training speakers was used as training material for feature detection for vowels. The detectors for the common features *vocalic*, *non-vocalic* and *voice bar* are trained using all the training material, that is three repetitions of burst-onset interval of stop consonants, one repetition on steady-state vowel, and one repetition of pre-voicing material from each of the three training speakers. The same arrangements with respect to input/output vectors for training and testing described in Section 6.2.1 applied to this experiment. The experiments were conducted on a CM-5 (from Thinking Machines Corporation, an MIMD — Multiple Instruction Multiple Data — Machine).

The results reported here used the same speaker for testing, for reasons of comparison with scheme A. Detailed results for the other speakers are reported in Sections 6.4 and 6.5.

Table 6.4 is a summary of the best architectures of the feature detectors and the correct detection rate of the features on the test speaker for the pseudo steady-state vowels. Table 6.5 reports the best architectures for the feature detectors and the correct detection rate of the features on the test speaker for the burst-onset intervals of stop consonants. Table 6.6 reports the results for common feature detection. The "best" architectures are those which gave the best performance on the test speaker (same speaker as in scheme A)(refer to Section 6.1.1 for architecture selection).

Feature	*Architecture*	*Rate %*
acute	13-6-2	97.7
compact	13-6-4-2	95.7
diffuse	13-12-2	96.4
flat	13-10-2	88.9
grave	13-8-2	94.2
lax	13-8-2	80.5
plain	13-4-2	87.8
tense	13-4-2	81.7

Table 6.4: *Scheme B:* Feature detection rates for pseudo steady-state vowels.

Feature	*Architecture*	*Rate %*
acute	13-10-2	92.9
compact	13-10-2	94.7
diffuse	13-8-2	95.0
grave	13-10-2	93.0
lax	13-10-2	65.0
tense	13-6-2	64.1

Table 6.5: *Scheme B:* Feature detection rates for burst-onset interval of stop consonants.

Feature	*Architecture*	*Rate %*
non-vocalic	13-12-2	98.4
vocalic	13-10-2	98.3
voice bar	13-8-2	99.3

Table 6.6: *Scheme B:* Feature detection rates for global Features.

Note that the results of detection for features *plain* and *flat* are not reported for the stop consonants, because by definition they have the same feature values for all the stop consonants (see Table 4.2) included in this study and they therefore do not serve to discriminate between stop consonants.

6.2.3 Conclusion

From the speech signal point of view, we assumed that the nature of the pseudo steady-state vowels and the burst-onset interval of the stop consonants are very different; i.e. the acoustic characteristics of the features for both types of signal are different. Although they share the same set of features in concept, they differ in their acoustic realisations. The experimental results demonstrate this point: feature detectors in scheme A (where the feature detectors are shared by both the burst-onset interval of the stop consonants and the pseudo steady-state vowels) gave worse results than the detectors in scheme B (where the burst interval of the stop consonants and the pseudo steady-state vowels have their own feature detectors). Given this result, scheme B was used for the feature detection level of our phoneme recognition system.

6.3 System Design

As described in Section 6.2, the realisation of the features in the stop consonants and in the vowels is different, therefore the recognition system was designed to have separate feature extraction modules for stop consonants and vowels. Ran and Millar (1993b) report some of the results described here.

As shown in Figure 6.3, the system is divided into three levels: a global feature detection level; a fine feature detection level; and a phoneme recognition level that combines the results of the feature detection levels.

The speech signal enters the system directly at both the first and second levels. The first level divides the speech data into three classes: *voice bar, vocalic* and *non-vocalic.* At the second level, fine features relevant to phoneme recognition of stop consonants for the *non-vocalic* classes and fine features relevant to phoneme recognition of the vowels for the *vocalic* classes are detected separately. At the final level, phoneme recognition decisions are made by combining the feature detection results.

The way in which the system works is as follows. For a given input, if the *Vocalic Detector* has its output class *on* "selected", it will activate the module *Fine Feature Detection for Vocalic Group.* The output of this module is used in order to assign the input signal to a vowel class. On the other hand, if the *Vocalic Detector* has its output class *off* "selected", it will activate the module *Fine Feature Detection for Non-Vocalic Group.* The output of this module is used in order to assign the input to a stop consonant class. This recognition decision can be enhanced by taking into account the output of the *voice bar* feature detector from the previous frames and by analysis of the transitional period from the stop consonant to the following vowel.

6.4 Global Feature Detection

The aim of the *Global Feature Detection* module is to separate the input speech signal into three groups: *voice bar, non-vocalic* and *vocalic.* Figure 6.4 illustrates how this module is composed of two sub-modules, each of which detects one global feature. The ("selected") output class *on* from *voice bar detector* sub-module activates the *Recognition of Stops* module. The ("selected") output class *on* from the sub-module *vocalic detector* activates the *Fine Feature Detection for Vocalic Group* module. The output class *off* from the sub-module *vocalic detector* activates the *Fine feature Detection for Non-vocalic Group* module. Sections 6.4.1 and 6.4.2 describe these modules in more detail.

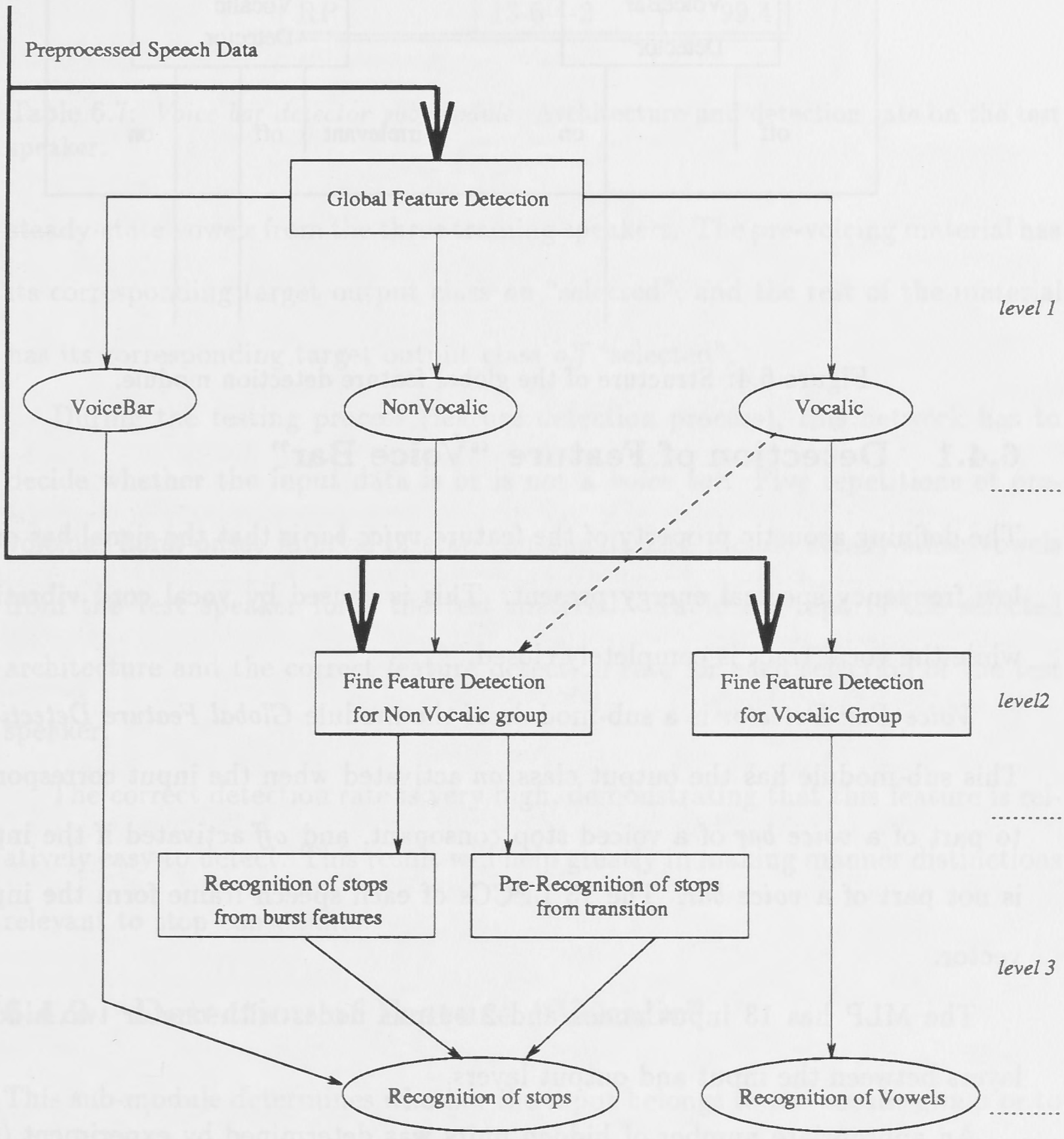


Figure 6.3: Overall system design.

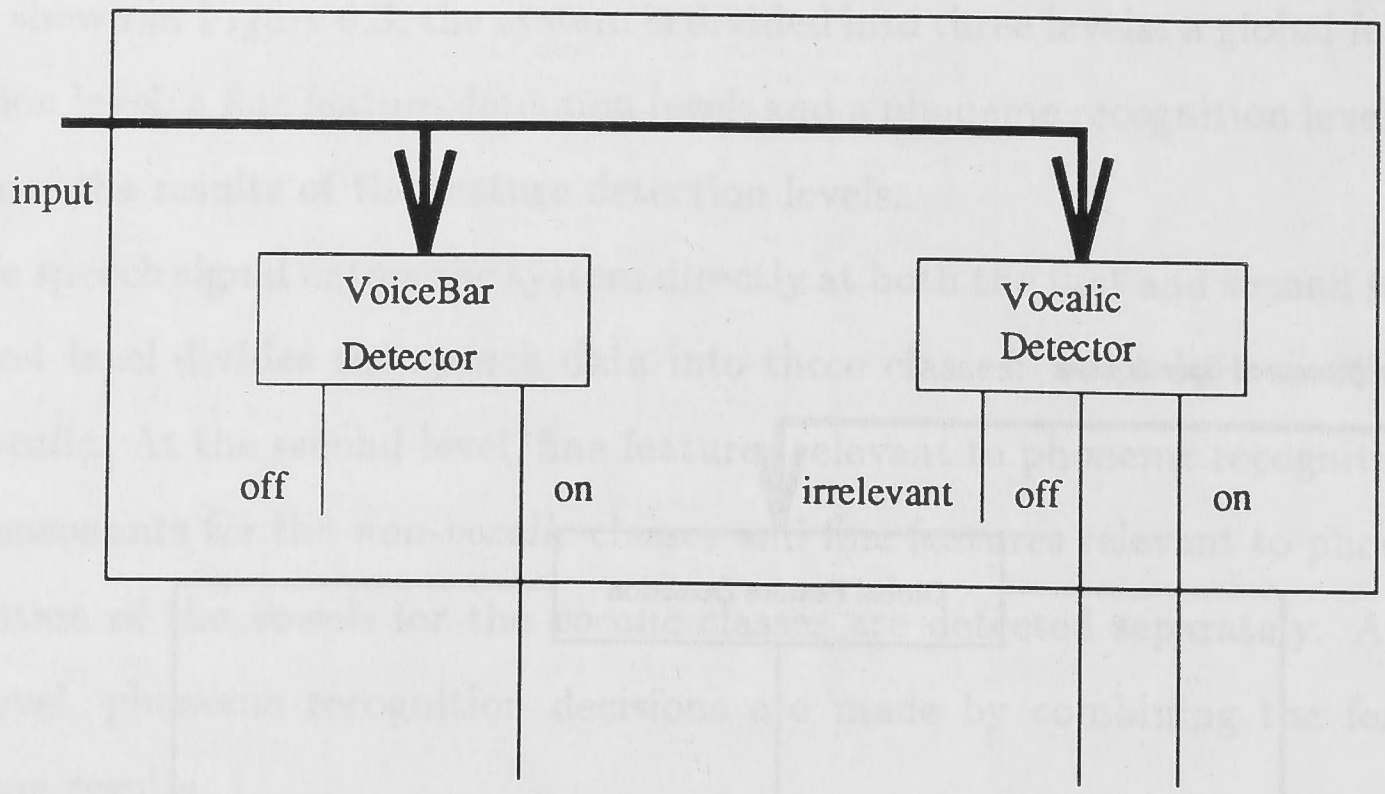


Figure 6.4: Structure of the global feature detection module.

6.4.1 Detection of Feature "Voice Bar"

The defining acoustic property of the feature *voice bar* is that the signal has only low frequency spectral energy present. This is caused by vocal cord vibration while the vocal tract is completely closed.

Voice Bar Detector is a sub-module of the module *Global Feature Detection.* This sub-module has the output class *on* activated when the input corresponds to part of a *voice bar* of a voiced stop consonant, and *off* activated if the input is not part of a *voice bar.* The 13 LPCCs of each speech frame form the input vector.

The MLP has 13 input nodes and 2 output nodes with one or two hidden layers between the input and output layers.

An appropriate number of hidden units was determined by experiment (see Section 6.1.1), choosing the trained ANN with best performance on the test data set. Table 6.7 summarises the architecture and the average correct detection rate of this sub-module for a complete rotation of test speakers. The "test speaker" column contains the test speaker's ID, the remaining three speakers being used as training speakers.

The training data comprises one repetition of pre-voicing material, three repetitions of burst-onset interval for stop consonants, and one repetition of pseudo

Test Speaker	*Architecture*	*Rate %*
IM	13-6-2	98.9
PM	13-8-2	99.3
PP	13-6-4-2	99.2
RP	13-6-4-2	99.4

Table 6.7: *Voice bar detector sub-module*: Architecture and detection rate on the test speaker.

steady-state vowels from the three training speakers. The pre-voicing material has its corresponding target output class *on* "selected", and the rest of the material has its corresponding target output class *off* "selected".

During the testing process (feature detection process), this network has to decide whether the input data is or is not a *voice bar*. Five repetitions of pre-voicing, burst-onset interval of stop consonants and pseudo steady-state vowels from the test speaker form the test material. Table 6.7 reports the selected architecture and the correct feature detection rate for each selection of the test speaker.

The correct detection rate is very high, demonstrating that this feature is relatively easy to detect. This result will help greatly in making manner distinctions relevant to stop consonants.

6.4.2 Detection of Feature "Vocalic"

This sub-module determines whether the input belongs to the *vocalic* group or to the *non-vocalic* group. By definition, the signal belongs to the *vocalic* group when there is a single periodic source whose onset is not abrupt. This feature serves to distinguish between the stop consonants and vowels, therefore, the pre-voicing speech material is regarded as being neither vocalic nor non-vocalic.

Figure 6.4 shows that the *Vocalic Detector* has 13 LPCCs as input and 3 output classes: *on, off, irrelevant.* The output class is *on* when the input belongs to *vocalic* group, *off* when the input belongs to *non-vocalic* group, and *irrelevant*

Test Speaker	*Architecture*	*Rate %*
IM	13-8-3	87.0
PM	13-10-3	98.3
PP	13-6-4-3	98.0
RP	13-6-3	98.3

Table 6.8: *Vocalic detector* sub-module: Architecture and detection rate on the test speaker.

when the input is not relevant to this feature. Specifically, if the input is burst-onset interval data, the output class should be *off*; if the input is from a pseudo steady-state vowel interval, the output class should be *on*; and if the input is from pre-voicing material, the output class should be *irrelevant*.

The following speech material was presented to the network during training: pre-voicing material with target class *irrelevant*, burst-onset intervals with target class *off*, and pseudo steady-state intervals of vowels with target class *on*. The testing phase corresponds to a feature detection process which decides whether the input belongs to one of the *vocalic* and *non-vocalic* groups, or was *irrelevant*. Five repetitions of all the material mentioned above from the test speaker was presented to the network. Table 6.8 reports the architecture and the percentage correct for detection of this feature. The CV transition was also tested and was classified as *vocalic* (i.e. with class *on* "selected"), except for the beginning few frames (two or three) in some cases. This is expected, because the CV transition which corresponds to the articulators moving from an initial stop consonant to the vowel target has a dynamic formant structure, and having a formant structure is characteristic of the *vocalic* feature. The beginning few frames of CV being classified as *non-vocalic* is due to the strong influence of the initial stop consonant, because these frames are immediately adjacent to the burst-onset interval.

6.5 Fine Feature Detection

At this level of the system (see Figure 6.3), the fine features for the *vocalic* and *non-vocalic* groups are detected, selecting those features which are useful in

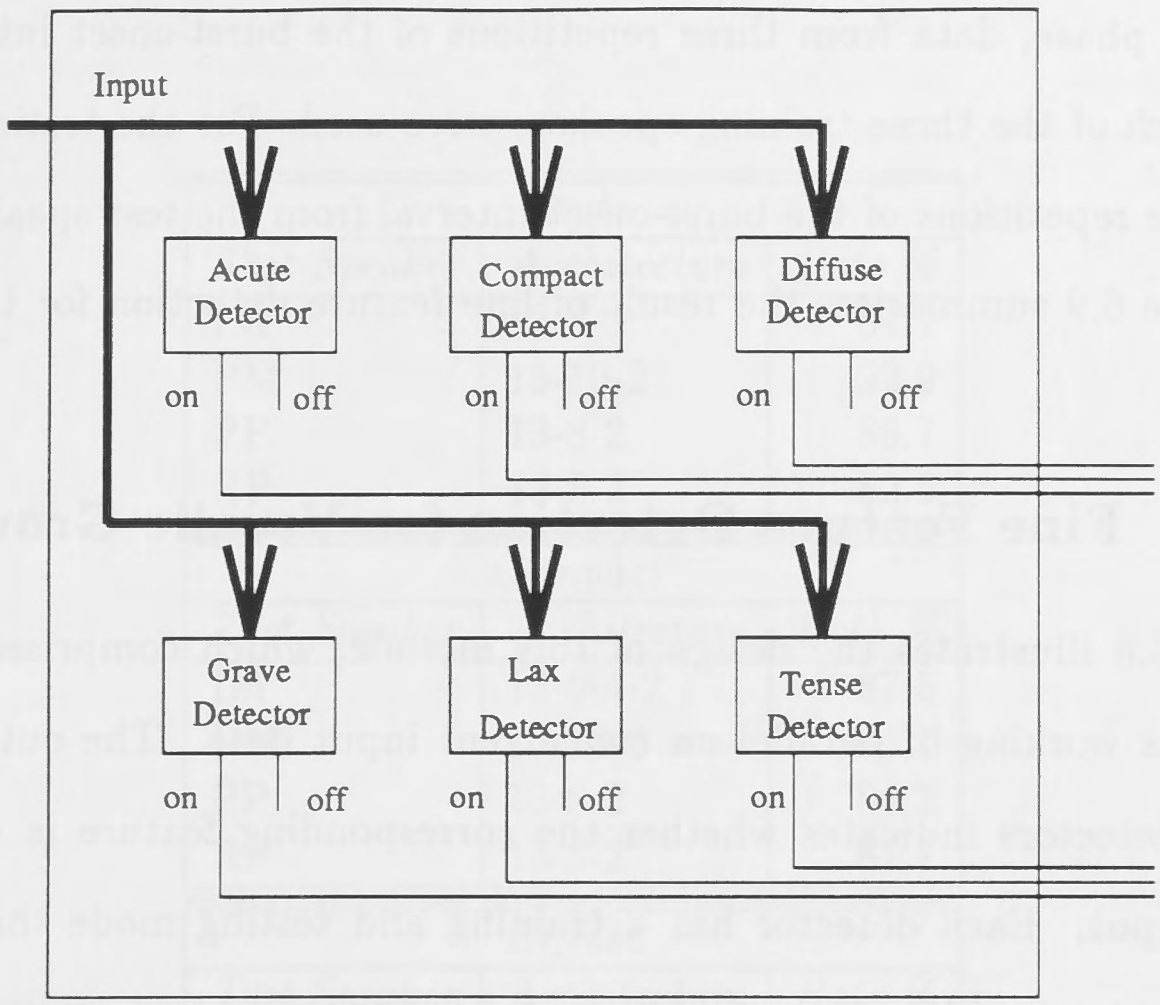


Figure 6.5: Structure of fine feature detection for *non-vocalic* group.

discriminating between the sounds in each group. For the *non-vocalic* group, the final goal of the system is to recognise the six stop consonants [**b**, **p**, **d**, **t**, **g**, **k**]. The features that are useful for discrimination of these stop consonants in the *non-vocalic* group are: *acute, compact, diffuse, grave, lax* and *tense*. For the *vocalic* group, the aim is to recognise the 11 monophthongs [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. The features that contribute to discrimination of these sounds are: *acute, compact, diffuse, flat, grave, lax, plain* and *tense*. There are two modules in this level, the *Fine Feature Detection for Non-Vocalic Group* and the *Fine Feature Detection for Vocalic Group* (as described in Sections 6.5.1 and 6.5.2).

6.5.1 Fine Feature Detection for Non-Vocalic Group

Figure 6.5 illustrates the design of this module, which consists of six feature detectors. Each feature detector is a sub-module, and is implemented by an MLP of specific architecture. The concurrent input to every sub-module is a frame of speech signal represented by 13 LPCCs.

A speech frame that is classified as belonging to the *non-vocalic* group by the *Global Feature Detection* module forms the input to this sub-module. For the

training phase, data from three repetitions of the burst-onset interval material from each of the three training speakers were used. For the testing phase, data from five repetitions of the burst-onset interval from the test speaker were used.

Table 6.9 summarises the result of fine feature detection for the *non-vocalic* group.

6.5.2 Fine Feature Detection for Vocalic Group

Figure 6.6 illustrates the design of this module, which comprises eight feature detectors working in parallel on concurrent input data. The output from each of the detectors indicates whether the corresponding feature is *on* or *off* for a given input. Each detector has a training and testing mode that corresponds to the feature evidence accumulation and feature detection process as described in Section 6.4.1. In the training phase, one repetition of pseudo steady-state vowels from three training speakers forms the training data. For the testing phase, pseudo steady-state vowels from five repetitions by the test speaker (whose data were not used during the training phase) was used. Table 6.10 reports the performance of the best architecture of the trained network.

6.5.3 Summary

In summary, fine feature detection accuracy is slightly lower than global feature detection accuracy (with a minimum of 98%, except in one case 87%). Fine feature detection accuracy is higher for the steady-state vowels than for the burst-onset intervals. The accuracy for vowels is above 95% except for features *tense/lax* and *flat/plain*. *Tense/lax* have detection accuracy above 81% and *flat/plain* have detection accuracy above 90%. The detection accuracy for burst-onset interval material is above 82% except for features *tense/lax* with a minimum of 64%. The low detection rates for burst-onset interval material and for *tense/lax* feature is most likely due to the duration related component of their nature. The current study models linguistic-phonetic knowledge — Jakobson *et al.*'s distinctive features which are *static* (see Sections 3.2.3.2 and 4.3). Under this static feature

Acute		
Test Speaker	*Architecture*	*Rate %*
IM	13-8-2	84.7
PM	13-10-2	92.9
PP	13-8-2	86.7
RP	13-8-2	82.6

Compact		
Test Speaker	*Architecture*	*Rate %*
IM	13-6-4-2	87.5
PM	13-10-2	94.7
PP	13-8-2	84.2
RP	13-8-2	81.7

Diffuse		
Test Speaker	*Architecture*	*Rate %*
IM	13-6-2	87.7
PM	13-8-2	95.0
PP	13-6-2	84.6
RP	13-4-2	81.5

Grave		
Test Speaker	*Architecture*	*Rate %*
IM	13-12-2	82.5
PM	13-10-2	93.0
PP	13-12-2	86.6
RP	13-8-2	82.6

Lax		
Test Speaker	*Architecture*	*Rate %*
IM	13-6-4-2	80.0
PM	13-10-2	65.0
PP	13-4-2	64.6
RP	13-4-2	64.5

Tense		
Test Speaker	*Architecture*	*Rate %*
IM	13-6-4-2	80.4
PM	13-6-2	64.1
PP	13-4-2	65.7
RP	13-6-4-2	63.5

Table 6.9: *Fine Feature detection for non-vocalic group module*: Architecture and detection rate on the test speaker.

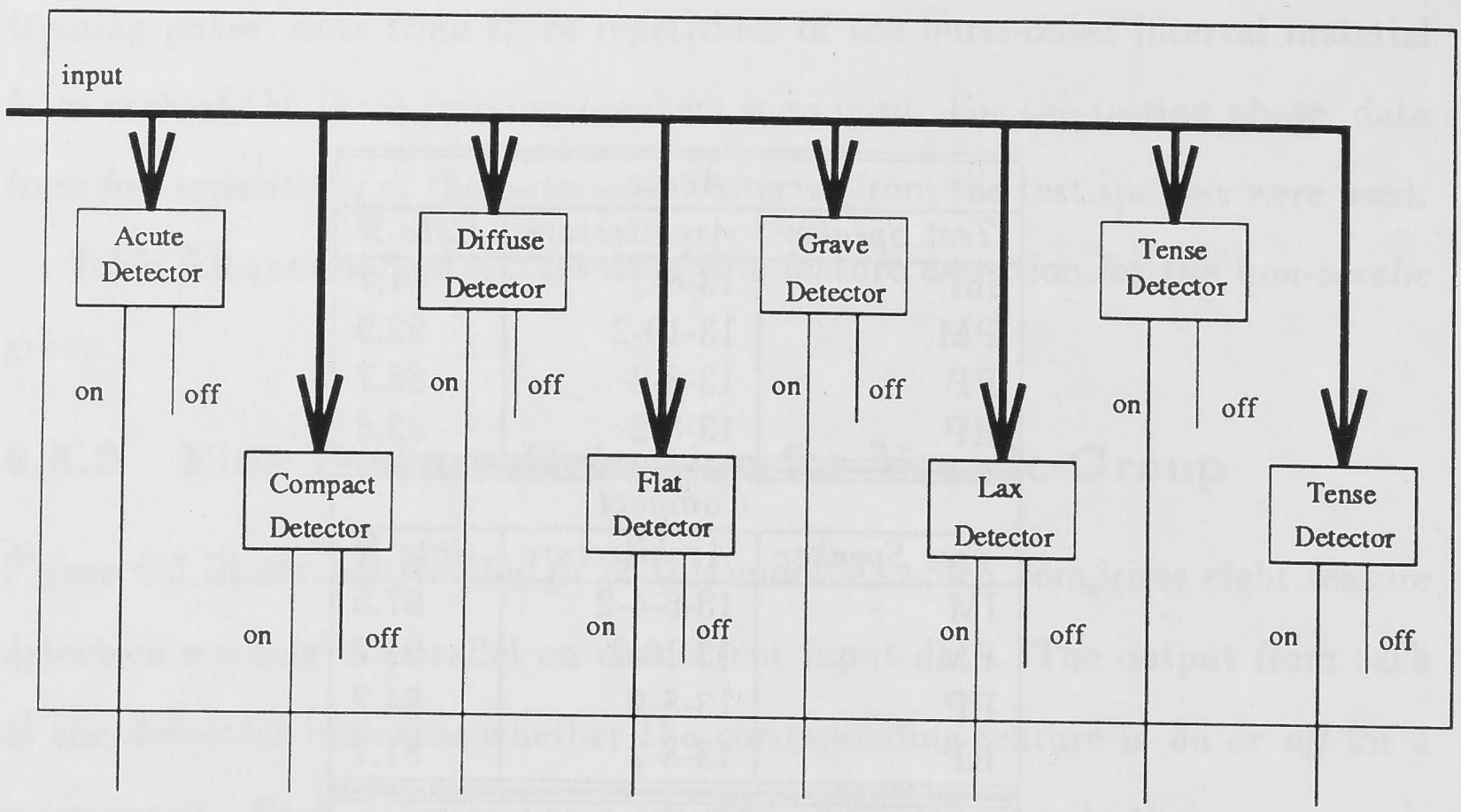


Figure 6.6: Structure of the fine feature detection for vocalic group.

framework, we choose the static parameters (LPCCs) to represent the signal. *Tense/lax* detection results demonstrate the limitation of this static framework, and highlight the need to modelling dynamic aspects of the signal. To improve *tense/lax* detection rates, the present *static* framework should be elaborated in order to incorporate dynamic information. Chapter 8 proposes a way forward here.

6.6 Phoneme Recognition Module

6.6.1 Introduction

The Phoneme recognition module aims to combine the fine feature detection results for each of the *non-vocalic* and *vocalic* groups in order to recognise the 6 stop consonants and the 11 monophthong vowels.

This section (*Phoneme Recognition Module*) describes an investigation into combining the fine feature detection results described in Section 6.5. Sections 6.6.2 and 6.6.3 describe recognition using (i) MLPs and (ii) table look-up techniques respectively.

Acute		
Test Speaker	*Architecture*	*Rate %*
IM	13-10-2	97.3
PM	13-6-2	97.7
PP	13-6-2	95.0
RP	13-4-2	98.0
Compact		
Test Speaker	*Architecture*	*Rate %*
IM	13-4-2	100.0
PM	13-6-4-2	95.7
PP	13-4-2	99.4
RP	13-4-2	97.2
Diffuse		
Test Speaker	*Architecture*	*Rate %*
IM	13-4-2	99.2
PM	13-12-2	96.4
PP	13-4-2	98.5
RP	13-8-2	96.9
Flat		
Test Speaker	*Architecture*	*Rate %*
IM	13-6-2	96.3
PM	13-10-2	88.9
PP	13-6-2	96.1
RP	13-4-2	93.7

Table 6.10: *Fine feature detection for vocalic group module*: Architecture and detection rate on the test speaker (*continued on page 108*).

Grave		
Test Speaker	*Architecture*	*Rate %*
IM	13-4-2	96.3
PM	13-8-2	94.1
PP	13-6-2	97.1
RP	13-6-4-2	91.5
Lax		
Test Speaker	*Architecture*	*Rate %*
IM	13-4-2	85.2
PM	13-8-2	80.5
PP	13-10-2	84.0
RP	13-6-4-2	89.3
Plain		
Test Speaker	*Architecture*	*Rate %*
IM	13-8-2	96.1
PM	13-4-2	87.8
PP	13-4-2	95.6
RP	13-4-2	95.0
Tense		
Test Speaker	*Architecture*	*Rate %*
IM	13-4-2	84.9
PM	13-4-2	81.7
PP	13-6-4-2	83.9
RP	13-4-2	90.2

Table 6.10: *Fine feature detection for vocalic group module*: Architecture and detection rate on the test speaker (*continued from page 107*).

6.6.2 Phoneme Recognition Using MLPs

This level of the system consists of modules for vowel recognition based on fine feature detection for the *vocalic* group and modules for stop consonant recognition based on fine feature detection for the *non-vocalic* group. These modules are also implemented by using MLPs. The process of training these modules is detailed in the next few sections. In the testing phase, a speech signal consisting of a series of frames representing each word is first pre-processed. The results of pre-processing consist of 13 LPCCs for each frame, passed on a frame by frame basis to the first and second levels of the system. The outputs of the relevant fine feature detection modules for the corresponding input are combined at this level, in order to derive the class of phoneme to which the input signal belongs. Sections 6.6.2.1 and 6.6.2.2 describe stop consonant and vowel recognition in more detail.

6.6.2.1 Stop Consonant Recognition

Three sub-modules from the previous levels contribute to stop consonant recognition: the sub-module for the *voice bar* feature detection; the sub-module that combines fine feature detection results of the *non-vocalic* group for *non-vocalic* data, and the sub-module that combines fine feature detection results from the *non-vocalic* group for the transitional part from the burst-onset interval to the pseudo steady-state vowel which is classified as belonging to *vocalic* group. The inclusion of this last sub-module is due to the fact that the preceding consonant's realisation is significantly affected by the following vowel. That is, the information contained by the transition where a stop consonant is released into a vowel also contains information about the consonant (e.g. Lieberman and Blumstein (1988), pp. 224-226).

(I) Stop Consonant Recognition from the Burst-Onset Interval

The aim of the experiments reported below is to investigate how well feature detection can be used for stop consonant recognition, and to analyse whether

it contributes to place distinction to the same extent as to manner distinction, although from the theoretical point of view, we expect that burst-onset interval contains mainly the place of articulation information of stop consonants.

We now focus on stop consonant recognition from the burst-onset interval by combining fine feature detection results for *non-vocalic* group data. In the recognition process, once the data is classified as *non-vocalic*, fine features in the *non-vocalic* group are then detected for this data at the second level of the system. At the third level (Figure 6.3), two groups of experiments are described: the recognition of six stops (including manner and place classification), and the recognition of place of articulation of the six stops (three classes: Labial ([**p**, **b**]), Alveolar ([**t**, **d**]), and Velar ([**k**, **g**])).

(I.a) Recognition of Six Stops

For the stop consonants, each of the features in each pair (*acute/grave, compact/diffuse, tense/lax*) has the same discrimination power, e.g. for a particular sound, if *acute* has value 0, *grave* should have value 1. Theoretically, therefore, to discriminate the six stop consonants, only one feature from each of the three pairs is required. In order to investigate the effect of adding redundant information to the recognition module, two experiments were conducted. One of them does not include the redundant information, using only three features ($n = 3$), whereas the other does, using six features ($n = 6$).

This sub-module is implemented by using one MLP that has n input nodes, and six output nodes classifying the input into six stop consonant classes, where n represents the number of features used as input in the experiment. There is one (or two) hidden layer(s) between input and output layers. Several experiments were conducted using different MLP architectures in order to select the best for each experiment, judged on the basis of the test speaker results (see Section 6.1.1). Experiments with input feature number n (three and six) were conducted. For $n = 3$, the features are selected one per pair from the following pairs of features:

Test Speaker	*Architecture*	*Rate %*
IM	8-12-6	60.9
PM	3-8-4-6	62.6
PP	3-6-6	47.5
RP	3-4-6	52.0

(a)

Test Speaker	*Architecture*	*Rate %*
IM	6-10-6	62.2
PM	6-6-6	62.1
PP	6-12-6	48.3
RP	6-6-4-6	52.3

(b)

Table 6.11: *Recognition of six stops from burst-onset interval material*: Architecture and correct detection rate on the test speaker: (a) With three input features; (b) With six input features.

acute/grave, compact/diffuse, lax/tense, by selecting the feature that has higher correct detection rate for the training speakers. For $n = 6$, the features are: *acute, compact, diffuse, grave, lax, tense.* The training data consist of the activation scores of the output layer from the fine feature detectors for *non-vocalic* group, where the input to these detectors is three repetitions of the burst-onset interval data of the three training speakers. Specifically, the selection of the three features is: for test speaker PM, *compact, grave, lax*; for IM *acute, diffuse, tense*; for PP *acute, diffuse, tense*; and for RP *compact, grave, tense.*

Table 6.11 reports the results of recognition of six stop consonants taking three and six features as input. The main errors occur in manner distinctions, i.e. between [**b**]-[**p**]; [**d**]-[**t**]; [**g**]-[**k**]. This is well expected, because from the theoretical point of view, the burst-onset interval mainly contains place of articulation information of stop consonants. The second group of errors is in place distinctions, i.e. between [**b**]-[**d**]; [**p**]-[**t**]; [**d**]-[**g**]; [**t**]-[**k**]. Generally speaking, the recognition accuracy improves by including more features, although it may not be significant.

(I.b) Recognition of Place of Articulation of Stop Consonants

The recognition of place of articulation of Australian English stop consonants includes Labial ([**b**, **p**]), Alveolar ([**d**, **t**]), and Velar ([**g**, **k**]). An experimental sub-module was implemented by using one MLP with n (number of input features) input nodes, one or two hidden layers, and one output layer of three nodes corresponding to the three classes of place of articulation. Three experiments were conducted by varying the number of input features n (2, 4, 6). The selection of features follows the same policy as described above in *Recognition of Six Stops* (Section 6.6.2.1(I.a)), that is to select one feature per pair of features which gave the best detection results. For $n = 2$, the input features are selected from *acute/grave, compact/diffuse* (one from each pair according to the performance of the feature detection on the training speakers). Specifically, for PM the two features are: *compact and grave*; for IM, *diffuse, acute*; for PP, *diffuse, acute*; for RP *compact, grave*. For $n = 4$, the input features are *acute, grave, compact, diffuse* for all rotations of test speakers. For $n = 6$, the input features are: *acute, grave, compact, diffuse, plain, lax, tense* for all rotations of test speakers. The training data are the activation scores from the output layer of the fine feature detectors for the *non-vocalic* group, whose input was three repetitions of the burst-onset interval from each of the training speakers.

Table 6.12 reports the results of place recognition of stop consonants taking two, four and six input features. The main errors occur in the distinction between labial vs. alveolar and alveolar vs. velar consonants. Comparing the results reported in Table 6.12(a) (for two input features), Table 6.12(b) (for four input features) and Table 6.12(c) (for six input features), we can conclude that the recognition result does not improve by including more features.

(I.c) Conclusion

Recognition of place of articulation of the six stop consonants was more accurate than recognition of the stop consonants themselves. This result is well expected

Test Speaker	*Architecture*	*Rate %*
IM	2-4-3	77.0
PM	2-6-3	93.0
PP	2-6-4-3	76.9
RP	2-4-3	78.9

(a)

Test Speaker	*Architecture*	*Rate %*
IM	4-6-4-3	76.1
PM	4-4-3	93.0
PP	4-4-3	77.4
RP	4-6-4-3	79.2

(b)

Test Speaker	*Architecture*	*Rate %*
IM	6-8-3	75.7
PM	6-6-4-3	92.3
PP	6-6-3	77.3
RP	6-6-3	79.0

(c)

Table 6.12: *Recognition of three places of articulation of stops from burst-onset interval material*: Architecture and correct detection rate on the test speaker: (a) With two input features; (b) With four input features; (c) With six input features.

from a theoretical point of view, because the burst-onset interval contains mainly the place of articulation information. The feature that determines the manner of articulation is the pair *lax/tense*. Table 6.9 shows that the correct detection rate for this pair of features is generally lower than for other features (by almost 20%). This illustrates one limitation of the static framework imposed by Jakobson *et al.*'s feature system. Improvement of *tense/lax* detectors taking dynamics into account will help to overcome this limitation (see Section 8.2.1). The Voice Onset Time can be one type of this dynamic information.

(II) Stop Consonant Recognition from the CV Transitional Period

In the <*stop*><*vowel*>[d] context, the initial CV transition corresponds to the articulators move from the initial stop consonant to the following vowel target.

Therefore, the CV transition is influenced by both the initial stop consonant and the vowel, where the beginning of the transition is more influenced by the stop consonant than by the vowel, and the final part of the CV transition is more influenced by the following vowel than by the stop consonant. The aim of this experiment is to look for spectral similarity to the burst in the start of the transition. To analyse experimentally how much stop information is contained in the CV transition period, several experiments were conducted by varying the number of frames of the CV transition included in the analysis, from one frame (i.e. 6.4 msec of the signal) up to five frames (i.e. 32 msec of the signal) starting from the beginning of the CV transition. The experiments were conducted for recognition of the six stop consonants and three places of articulation. Three input features were used.

(II.a) Recognition of Six Stop Consonants from CV Transition

One MLP is used for implementation of this sub-module. It consists of one input layer, one output layer and one or two hidden layers. The input layer has three nodes and the output layer has six nodes.

The training data correspond to activation scores from the output layer of the fine feature detection for the *non-vocalic* group whose corresponding input is three repetitions of the n frame CV transition from each of the three training speakers. Table 6.13 summarises the recognition results. The average recognition rate with test speakers were: using five frames immediately after the burst-onset interval – 50.1% (from Table 6.13(a)); using four frames – 52.0% (from Table 6.13(b)); using three frames – 54.3% (from Table 6.13(c)); using two frames – 55.0% (from Table 6.13(d)); using one frame – 57.3% (from Table 6.13(e)). This shows that a slight improvement results from inclusion of fewer frames. This leads to the conclusion that information adjacent to the of burst-onset interval contains the most relevant stop consonant related information. This is entirely predictable for a frame by frame analysis given the "noise" induced by the articulators move from

the stop consonant position to the next vowel target. Such "noise" representing systematic variations which are not however modelled satisfactorily in a static framework.

(II.b) Recognition of Place of Articulation of Stop Consonants from *CV* Transition

The same approach described in Section 6.6.2.1(II.a) (*Recognition of Six Stop Consonants from CV Transition*) is used for recognition of the three places of articulation: Labial ([**b**, **p**]), Alveolar ([**d**, **t**]), and Velar ([**g**, **k**]). Table 6.14 summarises the results.

The average recognition rates with varying numbers of frames were: five frames – 56.6% (from Table 6.14(a)); four frames – 58.8% (from Table 6.14(b)) ; three frames – 61.5% (from Table 6.14(c)); two frames – 64.7% (from Table 6.14(d)); one frame – 66.1% (from Table 6.14(e)). These results lead to the same conclusion as the previous section.

(II.c) Conclusion

The experiments confirmed our expectation that most consonant specific information is contained adjacent to the burst. In order to cope with dynamic changes, better models for *CV* transition regions are needed (see Section 8.2.2).

(III) Summary of Stop Consonant Recognition

The recognition result for place of articulation of stop consonants from the burst-onset interval material (from 75.7% to 93.0%, see Table 6.12) is much better than the result for place plus manner recognition (from 47.5% to 62.6%, see Table 6.11). This is expected from the theoretical point of view, because the burst-onset interval contains mainly place of articulation information. Combining the result of the feature *voice bar*, which also determines the manner of articulation of the stop consonants, with the result of place of articulation of stop consonants

Test Speaker	*Architecture*	*Rate %*
IM	3-8-6	56.6
PM	3-6-4-6	55.9
PP	3-8-6	46.7
RP	3-10-6	41.0

(a)

Test Speaker	*Architecture*	*Rate %*
IM	3-10-6	58.3
PM	3-8-6	58.0
PP	3-8-6	48.9
RP	3-6-4-6	42.6

(b)

Test Speaker	*Architecture*	*Rate %*
IM	3-10-6	59.6
PM	3-6-4-6	62.6
PP	3-10-6	50.4
RP	3-4-6	44.7

(c)

Test Speaker	*Architecture*	*Rate %*
IM	3-6-6	61.7
PM	3-6-4-6	60.9
PP	3-8-6	50.5
RP	3-4-6	47.0

(d)

Test Speaker	*Architecture*	*Rate %*
IM	3-8-6	61.2
PM	3-6-6	61.6
PP	3-10-6	50.9
RP	3-8-6	55.5

(e)

Table 6.13: *Recognition of six stops from CV transitional period*: Architecture and recognition rate on the test speaker: (a) From five frames of CV Transition; (b) From four frames of CV Transition; (c) From three frames of CV Transition; (d) From two frames of CV Transition; (e) From one frame of CV Transition.

Test Speaker	*Architecture*	*Rate %*
IM	3-6-3	60.2
PM	3-6-3	63.8
PP	3-6-4-3	57.4
RP	3-8-3	44.8

(a)

Test Speaker	*Architecture*	*Rate %*
IM	3-6-4-3	61.4
PM	3-10-3	68.1
PP	3-8-3	59.3
RP	3-6-3	46.3

(b)

Test Speaker	*Architecture*	*Rate %*
IM	3-4-3	62.4
PM	3-4-3	72.4
PP	3-10-3	62.6
RP	3-8-3	48.6

(c)

Test Speaker	*Architecture*	*Rate %*
IM	3-4-3	65.0
PM	3-4-3	76.0
PP	3-6-3	65.3
RP	3-8-3	52.4

(d)

Test Speaker	*Architecture*	*Rate %*
IM	3-4-3	67.0
PM	3-4-3	78.3
PP	3-8-3	70.3
RP	3-4-3	48.8

(e)

Table 6.14: *Recognition of three places of articulation of stops from CV transitional period*: Architecture and recognition rate on the test speaker: (a) From five frames of *CV* Transition; (b) From four frames of *CV* Transition; (c) From three frames of *CV* Transition; (d) From two frames of *CV* Transition; (e) From one frame of *CV* Transition.

from the burst-onset interval material, we expect to improve recognition results for manner and articulation to at least 75.7% which is the lowest recognition rate for place recognition (see Table 6.12). With better models for the features *tense/lax* and for the CV transition period incorporating dynamic information, we expect that the results can be significantly improved (from 47.5% to 62.6% to a better result than 75.7%).

6.6.2.2 Vowel Recognition

(I) Pseudo Steady-State Vowel Recognition

As described in Section 6.4.2, the *Vocalic* global feature detector responds to the burst-onset interval material as *off*, and to the rest of a syllable (including the pseudo steady-state vowel and the transition from the burst-onset interval to the pseudo steady-state vowel) as *on*. The data that is detected as *vocalic* will then go through the fine feature detectors for the *vocalic* group. The aim of the vowel recognition module is to combine these feature detection results in order to recognise the 11 monophthongs.

The vowel recognition module was implemented by using an MLP, and is trained on the fine feature detectors' outputs (i.e. the activation scores of the output nodes), where the input to the module is the material of one repetition of pseudo steady-state vowels from three training speakers. For each test speaker, the module is evaluated with the fine feature detectors' outputs, when the detectors inputs are from the pseudo steady-state vowel interval, or otherwise from the CV transition interval. The experiments were carried out by varying the number of features n (4, 6, 8) in order to analyse the effect of using redundant information as described above in Section 6.6.2.1(I.a) (*Recognition of Six Stops*). Each of the four features is selected from the four pairs: *acute/grave, compact/diffuse, lax/tense, and flat/plain.* For the test speaker PM, the four features are: *acute, diffuse, tense* and *lax.* For IM: *grave, compact, lax, flat.* For PP: *compact, flat, grave, lax.* For RP: *acute, compact, plain, tense.*

The six feature test included *acute, compact, diffuse, grave*, plus two features selected one from each of the two pairs, *lax/tense, flat/plain*; for PM: *tense, flat*; for IM: *lax, flat*; for PP: *lax, flat*; and for RP: *plain, tense.*

The eight feature test included all the features, namely *acute, grave, compact, diffuse, lax, tense, flat* and *plain.*

The vowel recognition results for the three groups are shown in Table 6.15 where Tables 6.15(a) to 6.15(c) indicate the best rate which corresponds to having four, six, and eight features as input respectively. Analysing the confusion matrix for each of the three cases (namely combining four, six and eight features), the errors occurred mainly (at least half of the total number of errors) in the following pairs: [æ]–[ɛ]; [i]–[ɪ]; [ɔ]–[ʊ]; [ɑ]–[ʌ]. These errors were expected, because of constraints of Jakobson *et al.*'s feature system (explained in the following). According to Table 4.3, /æ/ and /ɛ/ have the same feature definition, although the vowel height is different. Therefore, by definition, they cannot be discriminated under Jakobson *et al.*'s feature system. From Table 4.3, one can observe that these are three minimal pairs, i.e. they can be distinguished between themselves by only a single feature, i.e. *tense.* If this feature was detected correctly (assuming 100%) (see Section 6.5.3), the error rate for vowel recognition could be reduced significantly (at least by 50%, as more than half of the errors occurred are in the above error list) (e.g. in the case of including eight features as input). Comparing the results in Tables 6.15(a), (b) and (c), it is clear that better recognition results are obtained as more features are included as inputs in the experiment. So, in this case, it seems that redundancy helped to mitigate errors.

(II) Vowel recognition from the CV Transition

In the <*stop*><*vowel*>[*d*] context, the initial CV transition corresponds to the articulators move from the initial stop consonant to the following vowel target. Therefore, the CV transition is influenced by both the initial stop consonant

Test Speaker	*Architecture*	*Rate %*
IM	4-8-4-11	65.5
PM	4-8-11	75.9
PP	4-6-11	74.8
RP	4-10-11	85.8

(a)

Test Speaker	*Architecture*	*Rate %*
IM	6-8-11	84.1
PM	6-4-11	77.0
PP	6-12-11	81.9
RP	6-6-11	84.8

(b)

Test Speaker	*Architecture*	*Rate %*
IM	8-6-11	85.0
PM	8-8-11	77.0
PP	8-6-11	82.3
RP	8-10-11	85.3

(c)

Table 6.15: *Recognition of vowels from steady-state vowel material:* Architecture and recognition rates. (a) From four input features; (b) From six input features; (c) From eight input features.

and the vowel, where the beginning of the transition is more influenced by the stop consonant than by the vowel, and the final part of the CV transition is more influenced by the following vowel than by the stop consonant. We assume that the feature detectors trained on the pseudo steady-state vowels have formed their own acoustic criteria to judge whether the input signal has the acoustic properties corresponding to the features. We also assume that the CV transition near the vowel target shares these acoustic properties to a greater degree than the portion of the CV far from the target vowel, although the target vowel is not fully realised. The following experiment aimed to test this hypothesis within the vowel recognition framework, that is to detect features in the CV transition data using the detectors trained on the pseudo steady-state vowels. A recognition of the 11 vowels from the CV transition was then conducted on the basis of these feature detection results. The speech material for the CV transition of $<stop><vowel>[d]$ words was first passed through the feature detection module for the *vocalic* group. Recognition was performed on the features detected during the CV transition period.

Table 6.16 summarises the recognition results where the highest recognition rate is achieved with eight features as input. Some observations regarding these results are:

- The beginning of the CV transition tends to be misclassified. This is expected because the portion adjacent to the burst-onset interval has more stop consonant specific information (demonstrated in Section 6.6.2.1(II.a)).
- The greater the number of feature detectors used as input to the recognition, the higher is the correct rate of recognition (having four features as input performed worst, eight features was the best of all).
- Only moderate recognition performance was expected here. The purpose of this experiment was to test our hypothesis that the CV transition region shares the acoustic properties of the steady-state vowel features although

Test Speaker	*Rate (4 features)%*	*Rate (6 features)%*	*Rate (8 features)%*
IM	37.4	45.3	47.1
PM	4.9	32.7	39.4
PP	37.9	43.3	43.0
RP	3.1	15.4	52.4

Table 6.16: Recognition of vowels in CV transition.

not to the same degree as steady-state vowels. The part of the CV transition near the pseudo steady-state vowel shares these acoustic properties to a greater degree than the part which is far from it. The results of these experiments support our hypothesis. When modelled appropriately, information in the CV transition should contribute to better recognition accuracy when this information can be incorporated (see Section 8.2.2).

(III) Summary of Vowel Recognition Experiments

Table 4.3 shows that (for certain pairs of features) some vowel sounds do not have their feature value definition as one feature *on* and the other feature *off* (e.g. *acute* 'on' and *grave* 'off'). These sounds have both features as *off*. In this way Jakobson *et al.* represent more than two degrees of acuteness and compactness, corresponding to more than two degrees of frontness and openness (i.e. vowel height) as described in Section 4.3. The need to include both members of a pair of features (e.g. *acute* and *grave*) for better recognition results highlights the need to classify vowels as having more than two degrees of frontness and openness. Including both members of a pair of features is a practical way of implementing the theoretical point mentioned above. The experiments demonstrate that inclusion of more features for vowel recognition helps to cope with the non-binary nature of distinctive features — adding redundancy yields better results.

Another important point revealed from the experiments is that by explicitly detecting features prior to overall recognition, problems in the recognition can be analysed more easily. In this case, the major problem is due to the static framework constraint which leads to attempts to model the feature *tense/lax* in a static

framework (see Section 6.5.3), poor discrimination of this pair caused at least 50% of the recognition errors. This illustrates the benefit of using the approach adopted in this study, in which performance can be analysed, and problems can be diagnosed and corrected individually, without modifying other components of the system.

6.6.3 Phoneme Recognition Using Table Look–Up

Section 6.6.2 described the results of phoneme recognition using an MLP with input of continuous valued feature detection results. Phoneme classification is thus based on continuous feature values. Another possible approach is where the phoneme recogniser takes feature detectors' output as binary values, corresponding to a binary feature framework.

The present section describes experiments and results of phoneme recognition using such binary features. Comparison of this approach with the approach described in Section 6.6.2 should shed some light on the question of the binary/continuous nature of distinctive features while at the same time testing the impact of using different recognition techniques (i.e. MLP vs. table look-up).

The results of fine feature detection from the previous level of the system forms the input vector to this module. This vector contains n elements of value 0 or 1, with each element corresponding to one feature. The value is 1 if the corresponding feature is *on*, and 0 if the feature is *off*. This vector is then checked against a table which defines phonemes in terms of feature values as described in Tables 4.3 and 4.2. If the vector matches any phoneme's feature vector, then the corresponding frame is recognised as that phoneme, otherwise, the frame is labelled as unknown. This experiment is divided into stop consonant recognition and vowel recognition.

Test Speaker	*Rate %*
IM	53.9
PM	57.1
PP	47.5
RP	47.5

(a)

Test Speaker	*Rate %*
IM	48.5
PM	50.6
PP	42.1
RP	43.6

(b)

Table 6.17: Recognition of six stops using table look-up: (a) On three features; (b) On six features.

6.6.3.1 Stop Consonant Recognition

Stop consonant recognition using a table look-up technique was examined in two experiments: one based on use of three phonetic features and the other on use of six phonetic features. Selection of the three features was based on the same criteria described in Section 6.6.2.1, that is one feature from each of the pairs: *acute/grave, compact/diffuse, lax/tense* according to relative performance. The six features are all the features applicable to the stop consonants (*acute, grave, compact, diffuse, lax, tense*). Table 6.17 reports the results of recognising the six stop consonants for the case of three phonetic features and six features, using the table look-up technique described above. These are average results over five repetitions of the test speaker data.

The main errors occur in the manner distinction (i.e. /**b**/ recognised as /**p**/, /**d**/ recognised as /**t**/ and /**g**/ recognised as /**k**/), with other errors being made in the place distinction (e.g. /**p**/ as /**t**/; /**k**/ as /**t**/). More frames are labelled as unknown when more features are included in the recognition, because there are more possible uncategorised combinations of the feature values.

6.6.3.2 Vowel Recognition

A vowel recognition experiment using the table look-up technique was divided into three groups by having four, six and eight phonetic features as input, selected as described in Section 6.6.2.2. That is, the four features are selected one from each of the pairs: *acute/grave, compact/diffuse, lax/tense, flat/plain* according to the performance of the feature detectors on the training speakers. The six features were *acute, compact, diffuse, grave*, plus two features selected one from each of the pairs *lax/tense* and flat/plain. The eight features include all the features: *acute, compact, diffuse, grave, flat, plain, lax, tense.* Table 6.18 reports the results of recognition using table look-up. These results indicate that recognition rate is higher when use six features as input.

In the following, we list the main errors occurred. These errors are well expected, because of the constraints of Jakobson *et al.*'s feature system.

- /ɪ/, /i/: these sounds differ in duration represented by features *tense/lax.* Refer to Section 6.5.3 for an explanation.

- /ɛ/, /ɪ/: these sounds are front vowels differing in vowel height. This is represented by the features *compact/diffuse* where /ɛ/ has feature *–compact/+diffuse* and /ɪ/ has *–compact/–diffuse.* This is another example of the inadequate representation of vowel height by Jakobson *et al.*'s feature system, because it cannot distinguish the vowels which are neither high, nor low, and which have similar frontness but differ in vowel height.

- /æ/, /ɛ/: Table 4.3 indicates that there is no distinction between these vowels in terms of features. This is another case which shows an inadequacy of Jakobson *et al.*'s feature system.

- /ɑ/, /ʌ/: these differ in duration, as represented by *tense/lax.* Refer to Section 6.5.3 for an explanation.

Test Speaker	*Rate %*
IM	46.80
PM	61.13
PP	72.87
RP	62.45

(a)

Test Speaker	*Rate %*
IM	70.59
PM	60.12
PP	72.87
RP	67.33

(b)

Test Speaker	*Rate %*
IM	67.89
PM	49.78
PP	66.12
RP	64.10

(c)

Table 6.18: Recognition of vowels using table look-up: on (a) Four; (b) Six; and (c) Eight features.

- /ɔ/, /ʊ/: these differ in vowel height and duration (Refer to Section 6.5.3 for explanation).

- /u/, /ɔ/: these differ in vowel height and duration (see above comments regarding these differences).

6.6.3.3 Conclusion

Comparing the recognition results using the table look-up technique and the results using MLP, the latter are better than the former. Table look-up is a binary approach which makes decisions on the basis of binary features, whereas the MLP takes account of the continuous values of the distinctive features. These results suggest that the phonemes are realised by features of a continuous nature, as the recognition results are better when continuous values of the features are

used. These results provide evidence in support of the continuous nature of distinctive features, obtained in a speech recognition framework.

Analysing the errors which occurred, we discover that the main reasons for confusions are the following:

- Some errors occurred because of different vowel heights, which Jakobson *et al.*'s feature system does not represent. Therefore, this study demonstrates some limitations of Jakobson *et al.*'s feature system within speech recognition framework.

- Some errors occurred because of differences in duration, which is represented by the feature pair *tense/lax* (refer to Section 6.5.3 for an explanation and Chapter 8 for suggestions). This result shows that there is great advantage in using the approach proposed in this study, where by explicitly encoding distinctive features, we can relate the system performance to the system modules, facilitating the diagnosis and correction of problems. Improvements can be made to individual modules without modifying others.

6.7 Non-Feature Detection Approaches

For comparison reasons, this section describes two phoneme recognition experiments without explicit feature detection. One of them performs recognition for *vocalic* and *non-vocalic* groups separately. The other experiment performs recognition of the two groups of data together, which comprises the burst-onset interval and the pseudo steady-state vowel material. Section 6.8 compares and discusses the results described in this section and the results presented in Section 6.6.

6.7.1 Phoneme Recognition from Two Classes

This section reports results of phoneme recognition from two classes of speech material: *vocalic* and *non-vocalic*. The *non-vocalic* class contains speech material from burst-onset interval, and the *vocalic* class consists of speech material from

pseudo steady-state vowels. Recognition of the six stops was conducted on the *non-vocalic* group, and recognition of the 11 vowels was performed on the *vocalic* group. The input to the two sub-modules was 13 LPCCs per frame. Two sub-modules for this experiment are described in the following sections.

6.7.1.1 Stop Consonant Recognition

This module was implemented using an MLP. The training material consisted of three repetitions of burst-onset interval material from each of the three training speakers. Table 6.19(a) reports the recognition results for the six stops under this scheme. The recognition accuracies range from 54.9% to 74.8% with an average of 64.4%.

6.7.1.2 Place of Articulation of Stop Consonants Distinction

In order to make a complete comparison with recognition based on explicit features, an experiment in locating three places of articulation of stop consonants (Labial, Alveolar, Velar) was conducted. Table 6.19(b) reports the results of this experiment. The recognition accuracies range from 74.4% to 88.9% with an average of 81.9%.

6.7.1.3 Vowel Recognition

This sub-module was trained on the pseudo steady-state vowels of three training speakers. Table 6.19(c) reports the recognition results for the 11 vowels performed on the pseudo steady-state vowels of the test speaker. The recognition accuracies range from 67.5% to 84.6% with an average of 78.0%.

6.7.1.4 Vowel recognition from the CV Transition

To test how well the vowel recognition module under this scheme performs on the CV transitions, an experiment was conducted on the vowel recognition module described in Section 6.7.1.3 using speech material of CV transitions as the test data. Table 6.19(d) reports the results of this experiment. The recognition accuracies range from 37.3% to 44.4% with an average of 40.3%.

Test Speaker	*Architecture*	*Rate %*
IM	13-20-6	74.8
PM	13-20-6	63.8
PP	13-30-6	54.9
RP	13-30-6	64.1

(a)

Test Speaker	*Architecture*	*Rate %*
IM	13-20-3	87.4
PM	13-10-3	88.9
PP	13-10-3	76.8
RP	13-15-11	74.4

(b)

Test Speaker	*Architecture*	*Rate %*
IM	13-20-11	84.6
PM	13-20-11	67.5
PP	13-10-11	79.1
RP	13-20-11	80.9

(c)

Test Speaker	*Rate %*
IM	44.4
PM	37.3
PP	39.9
RP	39.5

(d)

Table 6.19: *Recognition under two monolithic groups scheme:* (a) Six stop consonants from the burst-onset interval material; (b) Three places of articulation of the stop consonants from the burst-onset interval material; (c) 11 vowels from pseudo steady-state vowel; (d) 11 vowels from the *CV* transitions.

6.7.2 Phoneme Recognition with A Monolithic Module on Burst and Vowel material

This section describes the result of a comparative experiment, where a single global module is employed. This module is trained on speech material from the burst-onset interval and pseudo steady-state vowel intervals. The training data were labelled as classes 1 to 17, where the first 11 classes correspond to the 11 vowel classes and classes 12 to 17 correspond to the 6 stop classes. The input was 13 LPCCs per frame.

Table 6.20(a) reports the best architecture found, and Table 6.20(b) reports the recognition results of these phonemes from the burst-onset interval, CV transition and steady-state vowel material.

The CV transitions were included as part of the testing material to see how well the module recognises vowels from CV transitions, having been trained on information from pseudo steady-state vowel segments only. Another reason was to compare the result with the counterpart approach, that is explicit feature detection prior to recognition.

Table 6.20(b) shows that the recognition accuracy for stops (from 42.7% to 54.1%) is lower than the recognition accuracy for vowels (from 58.5% to 66.8%) and the recognition accuracy of vowels from CV transitions is very low (from 24.5% to 35.7%). These results were expected because the CV transition is influenced by the stop consonant as well, therefore, the information learned by the MLP from training samples of vowels is not sufficient to characterise the transitions as vowels.

6.8 Summary and Discussion

To summarise and compare the results of all the experiments, Table 6.21 reports average recognition rates across test speakers for each experiment. Note that for clarity of presentation, the following notation is used throughout the rest of this chapter: *feature-MLP* represents the experiment in which the recognition

Test Speaker	*Architecture*
IM	13-20-17
PM	13-20-17
PP	13-35-17
RP	13-30-17

(a)

Test Speaker	*Rate (burst) %*	*Rate (vowel) %*	*Rate (transition) %*
IM	54.1	60.1	25.3
PM	47.0	61.6	28.3
PP	45.3	58.5	24.5
RP	42.7	66.8	35.7

(b)

Table 6.20: *Recognition under monolithic approach*: (a) Architecture; (b) Results in three groups for the Monolithic approach taking burst-onset interval and steady-state vowel material.

scheme used an MLP following feature detection; *feature-Table* represents the experiment in which table look-up techniques was used follow feature detection; *2stream* represents the experiment in which recognition of the six stop consonants was performed by a module dedicated to burst-onset interval speech material, and recognition of the 11 monophthongs was performed by a module dedicated to pseudo steady-state vowel speech material; *Mono* represents the experiment in which the burst-onset interval material and pseudo steady-state vowel material from the training material are directly input to a monolithic module to recognise the 17 phonemes.

Analysing the results expressed in Table 6.21, the following conclusions can be drawn:

- Under the explicit feature detection recognition scheme (*feature-MLP*, *feature-Table*), combination of the feature detection results using the MLP technique gives a better result than use of a table look-up technique (observe Table 6.21(a) and (c)). For stop consonant recognition (Table 6.21(a)), the recognition accuracy using an MLP after feature extraction is 55.8% and

feature-MLP		*feature-Table*		*2stream*	*Mono*
3feature	*6feature*	*3feature*	*6feature*		
55.8	56.2	51.5	46.2	64.4	47.3

(a)

feature-MLP			*2 stream*
2feature	*4feature*	*6feature*	
81.5	81.4	81.1	81.9

(b)

feature-MLP			*feature-table*			*2stream*	*Mono*
4f	*6f*	*8f*	*4f*	*6f*	*8f*		
75.5	82.0	82.4	60.8	67.7	62.0	78.0	61.7

(c)

feature-MLP			*2stream*	*Mono*
4feature	*6feature*	*8feature*		
20.8	34.2	45.5	40.3	28.7

(d)

Table 6.21: *Summary of recognition results:* (a) Six stop consonant recognition; (b) Place of articulation distinction for stop consonants; (c) 11 vowel recognition from pseudo steady-state vowels; (d) 11 vowel recognition from CV transitions.

56.2% using three and six input features respectively. Using table look-up, the accuracy is 51.5% and 46.2%. For vowel recognition (Table 6.21(c)), the accuracy using an MLP after feature extraction is 75.5%, 82.0% and 82.4% using four, six and eight input features respectively. Using a table look-up technique, the corresponding rates are 60.8%, 67.7% and 62.0% respectively. When using the table look-up technique, the input feature vector has values 1 or 0, denoting that the corresponding features are *on* or *off*. Binary *yes/no* decisions on the output of the feature detectors are required. There is no way to represent degrees of existence of a feature. By using an MLP technique, the activation scores of the output from the feature detectors are used as subsequent inputs, where the activation scores represent different degrees of the presence of features, or the probability of their existence. This more complete information is utilised by the MLP, giving better recognition results. These results demonstrate that binary features cannot discriminate the phonemes adequately, especially vowels, because Jakobson *et al.*'s feature system lacks to represent more than two degrees of frontness and openness. These experimental results strongly suggest that distinctive features are realised in a continuous manner acoustically.

- Comparing the results of two monolithic approaches (*2stream* and *Mono*) for the cases of stop consonant and vowel recognition, the scheme *2stream* gives the best result (Tables 6.21(a), (b) and (c) present results for stop consonants, vowels, and CV transitions respectively). The complexity of the information in the input to the MLPs increases from *2stream* to *Mono*, therefore the complexity of the problem given to the MLP increases. This result shows that the performance of MLPs can deteriorate as the input is made more complex and only the same amount of data is available for training. One solution to this problem is to break a big problem into smaller ones in order to gain better performance (as proposed in this study), where each module is dedicated to solving a small problem — feature detection.

Each module's task is to detect the presence of features from general properties of these features learned from training samples, not from the specific phoneme patterns presented in the training samples when using the monolithic approach. This modular approach facilitates generalisation.

- Table 6.21(d) shows the results of 11 vowel recognition from the CV transition material. The purpose of these experiments is to investigate how effective the pseudo steady-state models can be when used in the following two approaches: (i) recognition following explicit feature detection, (ii) monolithic recognition without feature detection.

  The results show that the approach with explicit feature extraction prior to recognition with eight input features is the best of all. This suggests that the feature detectors for vowels are extracting general properties of the features. When using this module to detect features from CV transitions, it outputs probabilities of presence of the features, which are subsequently combined by the MLP. In contrast, with *2stream* and *Mono*, the acoustic properties of the steady-state vowels are learned by the module during training. Obviously, the acoustic properties of CV transitions differ from those of the vowels, which leads to low recognition performance.

- Comparing the results of recognition using an MLP following explicit feature detection with the results of monolithic approaches, one can observe that the size of the best performing MLPs increases when the complexity of the problem given to the MLP increases. Specifically, in the case of recognition following explicit feature detection, 4 to 12 hidden nodes are used; in the case of *2stream*, MLPs having 10 to 30 hidden nodes are used; in the case of the monolithic approach *Mono*, MLPs having 20 to 35 hidden nodes are used. The training time for MLPs with more hidden units is much longer than for MLPs of less complicated architecture, if a suitable mapping can be found at all. It is also more difficult to diagnose and correct problems

in order to improve their performance. The number of samples for training a large MLP should increase with the size of the MLP, because there are more parameters to estimate from the training set.

- Table 6.21(a) summarises the results of recognition of the six stop consonants. It shows that the result of *2stream* (64.4%) is better than that of *feature-MLP* (up to 56.2%). This is partially due to (discussed in Section 6.6.2.1(I.a): *Recognition of Six Stops*) the static framework limitation which does not allow dynamic information relating to the feature pair *lax/tense* detectors to be modelled appropriately (see Section 6.5.3), which determines the manner of articulation. Table 6.21(b) summarises the results for place of articulation of the stops; this is an equivalent recognition result to that of six stops, assuming a perfect feature detector for *tense/lax*. The two schemes perform similarly (81.9% for *2stream* vs. 81.5% for *feature-MLP*).

- Table 6.21(c) shows that vowel recognition using explicit feature detection prior to recognition (with an MLP) outperforms monolithic approaches. With potential improvements in the detectors for the features *tense/lax*, the error rate of the former can be reduced by at least 50% (as discussed in Section 6.6.2.2), giving a recognition rate for the vowels of up to 91.2% (From Table 6.21(c), the maximum accuracy for *feature-MLP* is 82.4%; reducing errors by 50% the rate becomes 91.2%).

- This study concentrates on recognition of stops and vowels using the burst-onset interval and pseudo steady-state vowels respectively. A further study of the contribution to both consonant and vowel recognition from the CV transition, on the basis of explicit feature detection, is proposed in Section 8.2.2.

- Inclusion of more features for the recognition generally improves the recognition result, especially for the case of vowel recognition where the recognition accuracy increased about 7% on average by increasing the number

of features from four to six. In this case, the recognition accuracy did not increase significantly by increasing the number of features from six to eight.

In conclusion, recognition based on explicit feature detection prior to classification reduces the complexity of the architecture of the MLPs, and therefore improves the ability of the MLPs to find appropriate input/output mappings (with generalisation) given limited training data. By having such a modular organisation of MLPs, the feature detectors can be trained and used in parallel fashion, the intermediate results in the recognition process can be analysed easily, helping to diagnose and correct any problems, and independent improvements can easily be made to inadequate modules without the need to modify other modules. The recognition results are generally better than the monolithic approaches, especially for vowels.

One of the aims of this study is to investigate the practicality of explicit feature detection, where the features are linguistically oriented, and to see how such detection can be applied across speakers. This study demonstrates that it is feasible to detect a representative set of features, based on those identified in earlier research, on a speaker independent basis.

Another aim of this study is to investigate the utility of explicit feature detection for phoneme recognition. The results demonstrate that this approach can be applied successfully to vowel recognition. Elaboration of the distinctive feature framework to allow modelling of dynamic aspects should improve *tense/lax* detectors and other discriminations and lead to improved vowel recognition. The results also demonstrate that it is feasible to use feature detection prior to recognition for the stop consonants, although in this case too better models are needed to handle the dynamics of stop consonants and to improve recognition accuracy.

The experimental results strongly suggest that the acoustic realisation of distinctive feature is continuous in nature. This was demonstrated in a recognition framework where continuous features perform better than the binary features.

There is a continuing debate about the existence of invariant phonemic cues in the speech signal. Blumstein and Stevens (1979) insisted that such cues are present. They support their argument by extracting cues for place of articulation from the first 26 ms of stop consonants in CV context. They also showed that these cues are encoded in the gross shape of the spectrum sampled at the consonantal release. Their features are *diffuse-rising*, *diffuse-falling* and *compact* which correlate with Jakobson *et al.*'s features as *diffuse-rising* to *diffuse* and *acute*; *diffuse-falling* to *diffuse* and *grave*; and *compact* to *compact*. They reported an average accuracy in extracting these features of 85%. Our results, based on a relatively small amount of data, give average rates of extraction for features *acute*, *compact*, *diffuse* and *grave* in a nonvocalic context of 86.7%, 87%, 87.2% and 86.2% respectively. These results can be considered paralleling to those of Blumstein and Stevens with one difference: their results were based on the first 26 ms of stop consonants, whereas our results were obtained in a frame by frame analysis of stop consonants with frame lengths of 12.8 ms (6.4 ms overlap between adjacent frames) (see Chapter 5). Our results provide similar evidence of the existence of such invariant features. Further implications of our results for the theory of distinctive features are addressed in the following chapter.

Chapter 7

Analysis of the Feature Detectors

7.1 Introduction

Chapter 6 described phoneme recognition based on distinctive features in a modularly organised systematic way. This chapter considers certain fundamental issues regarding distinctive features and analyses the results of further experiments designed to clarify these issues.

The Figure 7.1 illustrates the vowels in Australian pronunciation (also presented in Figure 4.6) with the corresponding articulatory description labels and the distinctive features of concern. This clarifies the relationship between distinctive features and articulatory description of the vowels.

Following the definition of distinctive features, we label vowels in the top left corner of the figure as *acute* and *diffuse*, vowels in the bottom left corner as *acute* and *compact*, vowels in the top right corner as *grave* and *diffuse*, and vowels in the bottom right corner as *grave* and *compact*. The following question arises: how can one label intermediate vowels that are more centrally located? For example, the feature $+acute$ (or $-grave$) correlates with front vowels, and $-acute$ (or $+grave$) correlates with back vowels. How does one define central vowels in terms of *acute* (or *grave*), such as the vowel /ɜ/? Does /ɜ/ have the same degree of acuteness as /ɪ/, or the same degree of graveness as /ɔ/? Although theoretically some researchers prefer a binary approach for distinctive features (see Section 3.2.6), practically it seems logical to regard features as being present to different degrees in relation to the vowels (e.g. /ɪ/, /ɜ/, and /ɔ/ share the feature *acute* to different

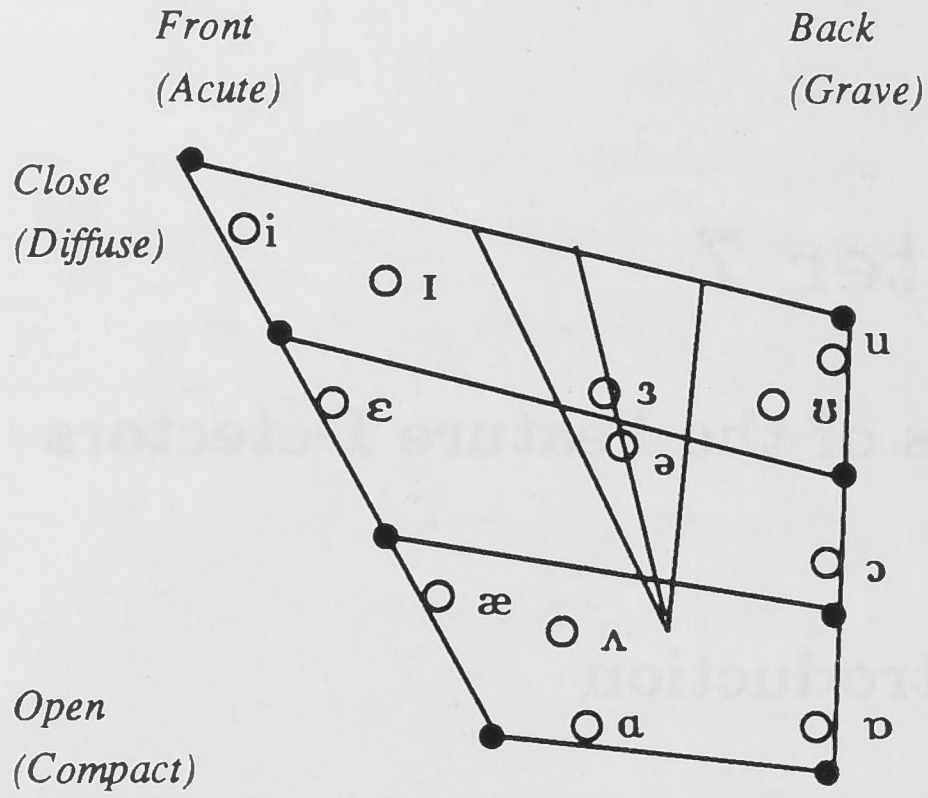


Figure 7.1: Vowels in Australian English (from Mitchell, 1962, p. 63) with corresponding labels of articulatory description and distinctive features.

degrees). If a distinctive feature is associated with a set of acoustic properties, the vowel sounds may share these acoustic properties to different degrees, despite the vowels having other acoustic properties that are distinct from this set. For the case of the feature *acute*, the associated acoustic property is energy concentration in the high frequency band. The different degrees may correlate with the position of the vowels in the articulatory space. To demonstrate this point, the following experiments were conducted. Evidence of certain features or of the acoustic properties of the features was extracted from those vowels that have a high degree of sharing of the features, and tests were performed on all 11 vowels. That is the training material consisted of three vowels and the test material consisted of the eight remaining vowels plus the training vowels (see Section 7.2 for details).

For the clarity and conciseness, some of the material on which this chapter is based has been placed in an appendix, on the following basis:

- We present results of the features *acute, compact* and *flat* in this chapter, and their counterpart features (i.e. *grave, diffuse* and *plain*) in Appendix A.

- When presenting results, we display figures from only one speaker to reduce the interruption to the text by the figures. Appendix A presents the complete set of results from all speakers. Discussion in this chapter is based on results from all speakers.
- This chapter includes material necessary for consideration to support the main issues; ancillary material is presented in the appendix.

Section 7.2 describes experiments which explore the extent of sharing of general acoustic properties, including training, testing and the method of analysis. Sections 7.3 and 7.4 analyse results for the features *acute* and *compact* respectively. Section 7.5 analyses results for feature *flat*. Analyses of results for features *grave*, *diffuse* and *plain* are presented in Appendices A.3, A.4 and A.6 respectively. In a further step, the results of the experiments for different features are combined to form a space analogous to the distinctive feature space (see Sections 7.6 and 7.7).

7.2 Design of the Experiments

The primary aim of these experiments is to investigate whether there exists a set of general acoustic properties relating to each distinctive feature whose components are shared by vowels to different degrees. To reduce the complexity of the problem, this study restricts the number of free parameters by eliminating the variability that is contributed by the speaker dimension, establishing a set of speaker dependent models (one model per distinctive feature) for each speaker. These models have training and testing phases as described in the following two sections.

7.2.1 Training

Classifier type Multi-Layer Perceptrons (MLPs) with a *sigmoid* activation function are used for the experiments described below, using a *back-propagation* algorithm with a *squared-error* cost function for training of the MLPs.

Each feature detector is implemented by using an MLP which has one input layer, one output layer and one hidden layer between them. The input layer consists of 13 nodes representing the 13 LPCCs. The output layer consists of two nodes representing *on* and *off* classes. The number of hidden nodes is determined by experiment.

As the aim of the experiments described in this chapter is to extract general properties concerning the distinctive features of concern from the reference vowel sounds, we aim to disregard the details of the vowel sounds. A broad principle of generalisation of MLPs is to use as simple as possible an architecture (Hush and Horne, (1993); also see Section 4.5.6) for the task in hand. Our training started with the least complex architecture possible, this coincided with the suggestions made by Hush and Horne (1993) on the selection of the MLP architectures to achieve good generalisation. In this case the starting architecture is 13-1-2. If the result when the model is tested with its training data is not satisfactory (i.e. $<$ 98%), the number of the hidden units is increased by 1. The final architectures for each of the detectors were found to be of size 13-2-2 or 13-3-2. The training process normally converged in less than 30 trials (i.e. presentations of the training data) for the finally selected architectures.

The input patterns come from the pseudo steady-state portion of three extreme vowels extracted from each of the six stop consonant environments. These extreme vowels are ɪ, ɒ, and ʊ. The term *extreme* notes that they are relatively stable vowels near the corners of the cardinal vowel system. They share the acoustic properties of the features to a great degree. ɪ is a *front-close unrounded* vowel, ɒ is a *back-open unrounded* vowel and ʊ is a *back-close rounded* vowel. These vowels can be well defined in terms of the phonetic features analysed herein, as defined in Table 7.1. Each output target vector has one element set to one and the others to zero (see Section 4.5.5).

7.2.2 Testing

Tests were performed on all 11 vowels; results are given in Sections 7.3 through 7.7. The test results can be interpreted in terms of the probability of a given sample being classified as having the feature *on* (see Section 4.5.5).

The trained MLP acts according to the extracted general criteria from the training samples. For the case of the *acute* feature, the corresponding detector may look for an energy concentration in the high frequency band, according to the energy in the high band of a given sample. It responds with a confidence score (probability) indicating how well the given sample could be classified as having the feature (e.g. +*acute*). A low energy in the high frequency band relates to a low output confidence score.

In summary, despite the fact that vowels have different acoustic properties and given that the trained MLP forms its own criteria from the general properties extracted from the training samples relating to a given feature during training, during testing the MLP responds in the general properties' domain. Therefore, although vowels have other acoustic properties, only the relevant acoustic properties are tested within the domain of the feature's general properties.

7.2.3 Method of Analysis

When testing a trained detector with a given test pattern, the output from each output node is a continuous value [0..1]. This value is the *activation score* of the corresponding output node which can be used as the *degree of confidence* of the input pattern being classified in a particular feature class or the probability that the particular feature is present in the input pattern (see Section 4.5.5).

For each of the features analysed in the following few sections, the activation scores from a particular feature detector for all the vowels are graphed on the same axes, where each point represents the average of the activation scores of all the frames of the same steady-state vowel. Each average activation score indicates the degree of confidence of having the input vowel x assigned to the output class

Features	*Vowels*		
	ɪ	ɒ	ʊ
Vocalic/Non-Vocalic	+	+	+
Compact/Diffuse	−	+	−
Grave/Acute	−	+	+
Flat/Plain	−	−	+
Tense/Lax	−	−	−
Voice Bar	−	−	−

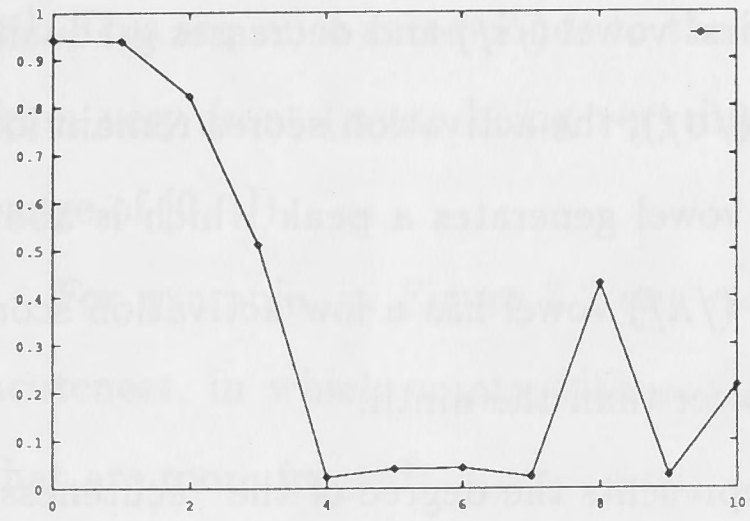
Table 7.1: *Feature Values for Vowels:* in the *(x/y) pair of features, "+" means feature x is on and y is off; "−" means feature x is off and y is on.*

on for a given feature. In other words, it represents how likely the feature is to be present in the input vowel x. The order of the points presented in the associated graphs is the same order as that of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. The horizontal axis enumerates the vowels; the vertical axis represents activation scores ranging from 0.0 to 1.0.

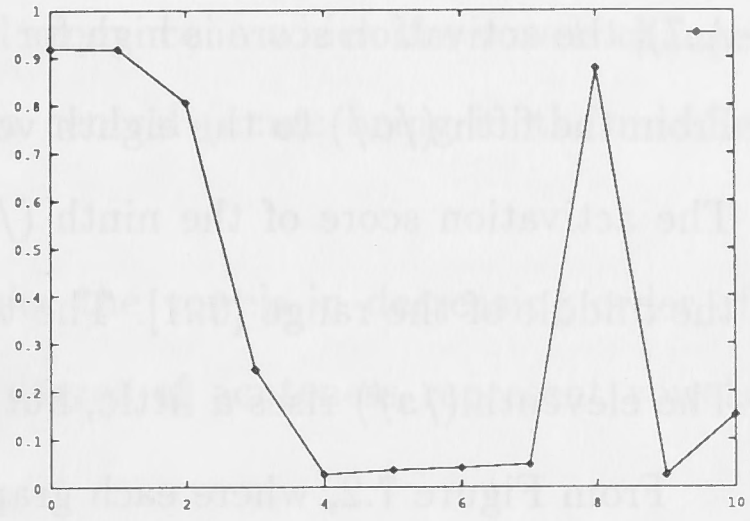
7.3 Analysis of the Feature "Acute"

Each speaker has a detector for the feature *acute* whose architecture is 13-3-2. The training material comprises one repetition of three pseudo steady-state reference vowels [ɪ, ɒ, ʊ] from one speaker in all six consonantal contexts.

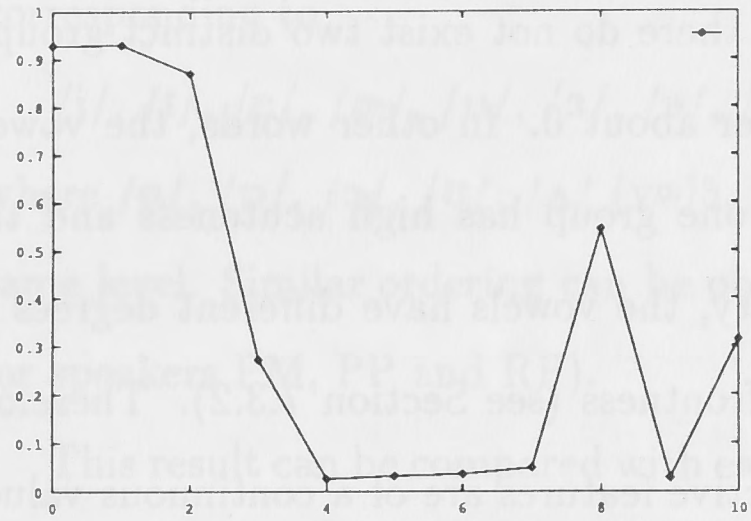
In the testing process, speaker dependent tests and speaker independent tests are performed. Figure 7.2 reports the test result of speaker IM's *acute* feature detector, using speaker IM's data. In this figure, there are six graphs, where each graph presents the result of 11 vowels in the context of a stop consonant. In the same way, Figures A.2, A.3 and A.4 report test results for the *acute* feature detectors trained and tested using data from speakers PM, PP and RP respectively. Figures A.5, A.6, and A.7 report test results of *acute* feature detectors for speakers IM, PP and RP tested on speaker PM's data. The next few sections discuss these results. Results of the related feature *grave* are analysed in Appendix A.4.



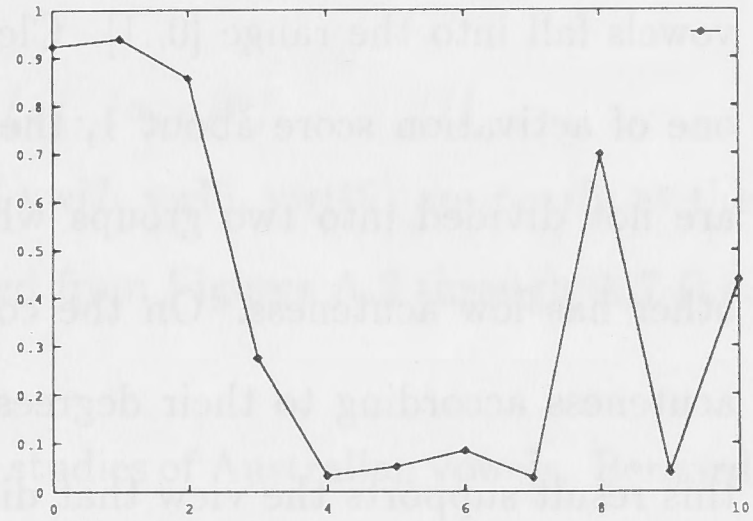
(a)



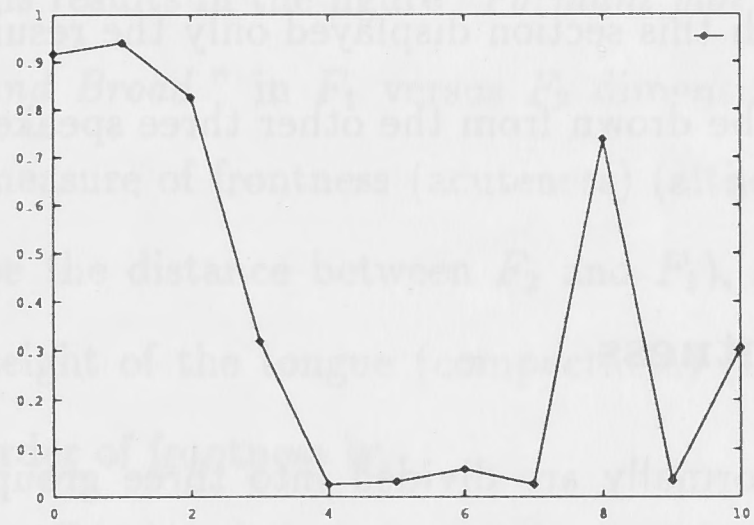
(b)



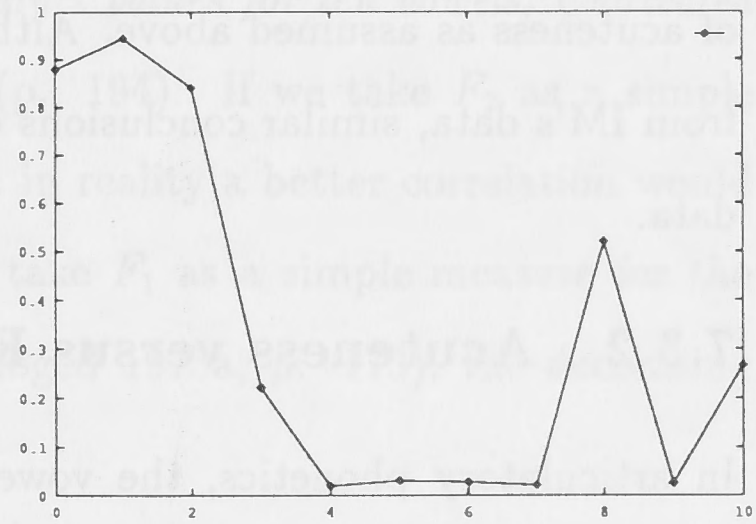
(c)



(d)



(e)



(f)

Figure 7.2: *Test results of IM's* acute *feature detector tested on IM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ] (labelled 0 to 10, refered as first to eleventh vowel. For example the vowel with label 5 in the graph is the vowel ɒ, and is refered to as the sixth vowel; the vowel with label 9 in the graph is the vowel ʌ, and is refered to as the tenth vowel.).

7.3.1 Binary versus Continuous

Observe Figure 7.2 (similar observations can be made from Figures A.2 through A.7); the activation score is high for the first vowel (/i/) and decreases gradually. From the fifth (/ɑ/) to the eighth vowel (/ʊ/), the activation scores remain low. The activation score of the ninth (/u/) vowel generates a peak which is about the middle of the range [0..1]. The tenth (/ʌ/) vowel has a low activation score. The eleventh (/ɜ/) rises a little, but is lower than the ninth.

From Figure 7.2, where each graph represents the degree of the "acuteness"[1] of each of the 11 monophthongs, one can observe that the activation scores of the vowels fall into the range [0..1]. Clearly there do not exist two distinct groups, one of activation score about 1, the other about 0. In other words, the vowels are not divided into two groups where one group has high acuteness and the other has low acuteness. On the contrary, the vowels have different degrees of acuteness according to their degrees of frontness (see Section 7.3.2). Therefore this result supports the view that distinctive features are of a continuous valued nature (Fant, 1973, p. 152), given that /ɪ/, /ɒ/, and /ʊ/ represent the extremes of acuteness as assumed above. Although this section displayed only the results from IM's data, similar conclusions can be drawn from the other three speakers data.

7.3.2 Acuteness versus Frontness

In articulatory phonetics, the vowels normally are divided into three groups: front, central and back vowels corresponding to the highest point of the tongue being at the front, middle or back of the mouth respectively (Ladefoged, 1982, p. 12). Frontness is thus an attribute that describes how far front is the constriction formed by the tongue in articulating a vowel. The feature *+acute* (or *−grave*) represents front vowels while *−acute* (or *+grave*) represents the back vowels (see Figure 7.1 and Section 4.3). Therefore, acuteness correlates with frontness.

[1] by "acuteness" I mean the MLP output after it has been trained to make "acuteness" distinctions.

From Figure 7.2, it can be seen that there is a gradual decrease from the second vowel (/ɪ/) to the fifth vowel (/ɑ/). The second to fifth vowels are [i, ɪ, ɛ, æ]. This sequence corresponds to the vowel sequence in which the vowels change from very front (*acute* being very high) to central (*acute* being in the middle range of [0..1]).

For example, in Figure 7.2 one can order the vowels in decreasing order of acuteness, in which points with a higher degree of acuteness represent vowels that are more fronted:

vwl0, vwl1, vwl2, vwl3, vwl8, vwl10, vwl6, vwl5, vwl9, vwl4, vwl7

corresponding to

/i/, /ɪ/, /ɛ/, /æ/, /u/, /ɜ/, /ɔ/, /ɒ/, /ʌ/, /ɑ/, /ʊ/, *(1)*

where /ɑ/, /ɒ/, /ɔ/, /ʊ/, /ʌ/ (vwl5, vwl6 vwl7, vwl8, vwl10) are nearly at the same level. Similar ordering can be obtained from Figures A.2 through A.7 (i.e. for speakers PM, PP and RP).

This result can be compared with earlier studies of Australian vowels. Bernard (1989), reports his findings of an extensive study of Australian vowels. He plotted his results in the figure "*Formant plot of target values for ten vowels, Cultivated and Broad* " in F_1 versus F_2 dimensions (p. 194). If we take F_2 as a simple measure of frontness (acuteness) (although in reality a better correlation would be the distance between F_2 and F_1), and take F_1 as a simple measure for the height of the tongue (compactness) (Ladefoged 1975a, p. 173), the decreasing order of frontness is:

/i/, /ɪ/, /ɛ/, /æ/, /u/, /ɜ/, /ʌ/, /ɑ/, /ɒ/, /ʊ/, /ɔ/. *(2)*

The decreasing order of frontness of vowels from Mitchell's vowel chart (Mitchell, 1962, p. 63; Mitchell 1946, p. 30; Mitchell and Delbridge 1965, p. 35) is:

/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /ɜ/, /ʊ/, /u/, /ɔ/, /ɒ/, *(3)*

where /u/, /ɔ/, /ɒ/ are nearly at the same level.

Comparing the result of the present study with Bernard's and Mitchell's findings, the front and the central vowels correlate with Bernard's findings, and the back vowels are more like Mitchell's.

These results (from all the speakers) demonstrate that the order of vowels by their acuteness (response from the *acute* feature detector) is similar to the vowels ordered according to their frontness reported in the literature (compare orders of vowels presented in sequences *(1)*, *(2)* and *(3)*). This result indicates that the *acute* feature detectors provide relative frontness measures.

One observation from the sequences of vowels is: *(1)* and *(2)* have the vowel /**u**/ in the fifth place, in other words, the fifth most fronted vowel which means this vowel is a central vowel. In the sequence *(3)*, the vowel /**u**/ is the ninth which means /**u**/ is a back vowel. It is therefore clear that our results regarding to the vowel /**u**/ are in agreement with Bernard's, and not with Mitchell's (also see Figure 7.1). In conclusion, our *acute* feature detection results are consistent with Bernard's acoustic analysis result regarding the vowel /**u**/, that is /**u**/ is centralised vowel in Australian English (see Section 7.3.4 for more details).

7.3.3 Acuteness versus Acoustic Parameters

Ladefoged (1975a, p. 173) indicates that $F_2 - F_1$ correlates with frontness and F_1 correlates with the height of the tongue which characterises openness. Section 7.3.2 described our results regarding acuteness and compared these with other studies in terms of frontness. This section analyses our acuteness results by relating them directly to acoustic parameters, that is to $F_2 - F_1$. Table 7.2 summarises the values of F_1, F_2 and $F_2 - F_1$ of the 11 vowels in context of [**bVd**] for the speaker IM.

To relate acuteness (which correlates to frontness in articulatory terms) directly to $F_2 - F_1$ (which characterises frontness in acoustic terms) (Ladefoged, 1965a, p. 173), Figure 7.3 presents a graph of $F_2 - F_1$ versus acuteness, where the horizontal axis represents $F_2 - F_1$ and the vertical axis represents acuteness in terms of the activation scores of the *acute* feature detector, and each point represents a vowel. The conclusion from this figure is that the relationship between $F_2 - F_1$ and acuteness is close to linear in the region between the reference vowels.

Vowel	F_1 *(Hz)*	F_2 *(Hz)*	$F_2 - F_1$ *(Hz)*
/i/	352	2313	1961
/ɪ/	408	2148	1740
/ɛ/	501	1990	1489
/æ/	617	1835	1218
/ɑ/	752	1250	498
/ɒ/	610	938	328
/ɔ/	468	797	329
/ʊ/	408	921	513
/u/	380	1625	1245
/ʌ/	708	1323	615
/ɜ/	509	1464	955

Table 7.2: F_1, F_2, and $F_2 - F_1$ for 11 vowels in the context [bVd] from the speaker IM.

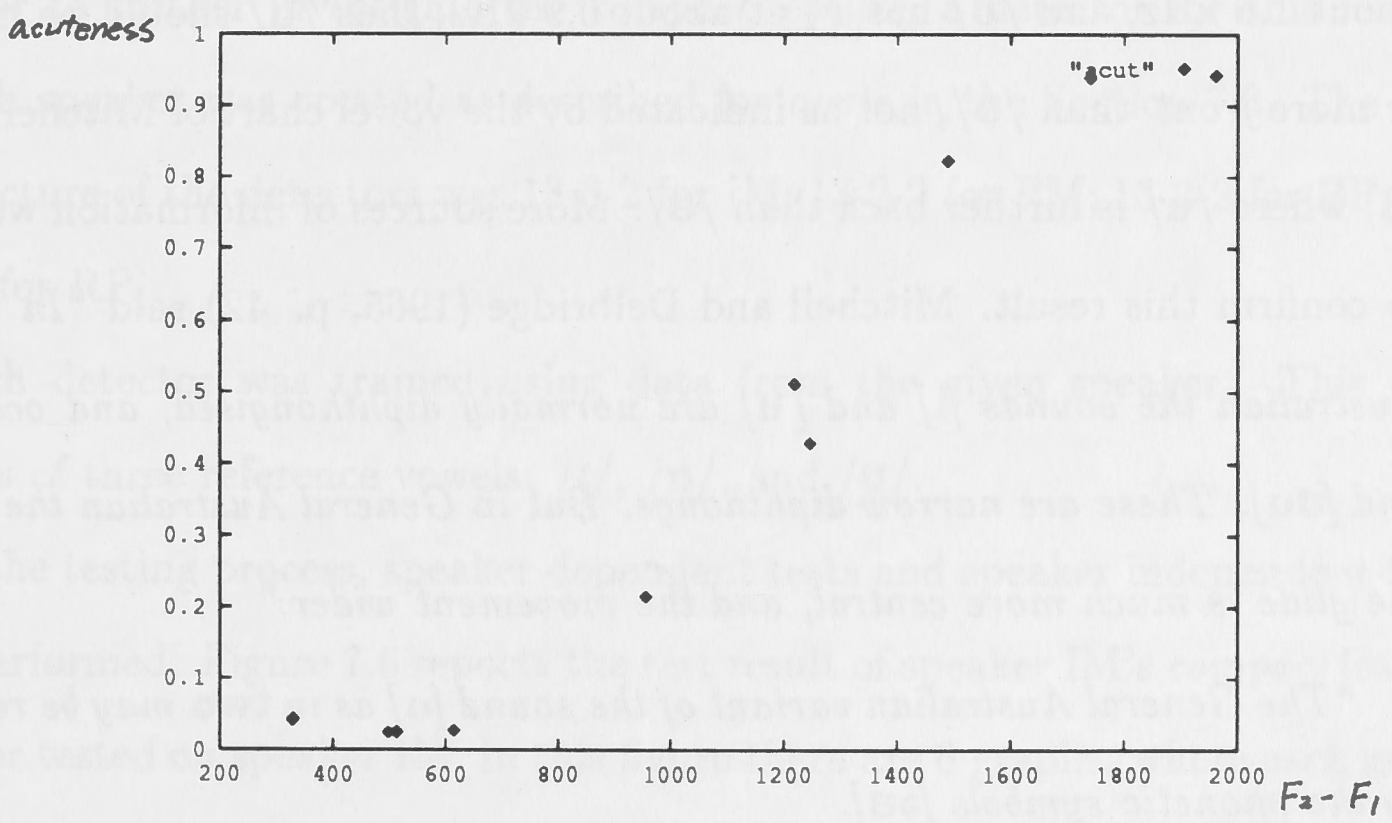


Figure 7.3: Relationship between $F_2 - F_1$ and acuteness based on speaker IM in [bVd] context.

7.3.4 The Vowel /u/

Observing Figure 7.2, one can notice that the vowel /**u**/ is the fourth or fifth most *acute* of the 11 monophthongs, therefore /**u**/ is the fourth or fifth most fronted vowel. This finding is totally different from Mitchell (1962) (see Figure 7.1, where /**u**/ is further back), on which this thesis was originally based. Mitchell (1962) described vowels in Australian pronunciation with reference to cardinal vowels (see Figure 4.6, extracted from Mitchell 1962, p. 63, Figure 21). In the description of the vowel /**u**/ based on an Educated (or Cultivated)[2] speaker of Australian English, Mitchell (1962, p. 69) said "*In the pronunciation of most people* **u** *is not precisely a vowel but a slight diphthong in which the tongue begins in the position for /***ʊ***/ and moves towards /***u***/. If we want to be quite precise we may represent the sound as* **ʊu**".

To analyse the results of this study, and to account for the differences between these results and Mitchell's, spectrograms of /**ʊ**/ in the context [**bʊd**] and /**u**/ in the context [**bud**] may be compared (see Figures 7.4 and 7.5). /**u**/ has F_2 at about 1.6 kHz, and /**ʊ**/ has F_2 at about 0.9 kHz, thus /**u**/ should be more *acute* or more *front* than /**ʊ**/, not as indicated by the vowel chart of Mitchell (1962, p. 63) where /**u**/ is further back than /**ʊ**/. More sources of information were sought to confirm this result. Mitchell and Delbridge (1965, p. 42) said "*In Cultivated Australian the sounds [***i***] and [***u***] are normally diphthongised, and occur as [***ɪi***] and [***ʊu***]. These are narrow diphthongs. But in General Australian the quality of the glide is much more central, and the movement wider.*"

"*The General Australian variant of the sound [***u***] as in* **two** *may be represented by the phonetic symbols [***əʊ***].*

---

[2]According to Mitchell and Delbridge (1965), there are three varieties of Australian English: Broad, General and Cultivated. The classification is based on vowel quality. The characteristics of Broad Australian English are: an undeveloped type of utterance lacking in fluency and characterised to a greater degree than the other varieties by excessive nasality, fusion in the segments, and a lack of clear structure in the prosodic features of stress and pitch. The type General represents the majority of Australian speakers and is between Broad and Cultivated.

The tongue begins at a position on the central vowel line and about half-way between half-close and close, and moves towards a position almost half-way from central to back and about a quarter of the distance from half-close to close."

Furthermore, our results were compared with Bernard's (1989) findings regarding /**u**/ and were found to be in agreement for both Cultivated and Broad (see Section 7.3.4 for definition of these classes) where /**u**/ is more centralised.

In conclusion, the experimental results of this study show that /**u**/ is centralised in the four analysed speakers. What appeared to be an anomaly when first compared with Mitchell (1962), is shown to be in line with the acoustic facts and in agreement with Bernard (1989) and Mitchell and Delbridge (1965). The discrepancy between our results and Mitchell (1962) regarding the vowel /**u**/ may be due to the fact that the four analysed speakers are of type General rather than Cultivated, in the strict sense imposed by Mitchell (1962), and Mitchell and Delbridge (1965).

7.4 Analysis of Feature "Compact"

In order to further investigate the *compact* feature, a detector for this feature for each speaker was created as described for *acute* in the Section 7.3. The best architecture of the detectors was 13-3-2 for IM; 13-2-2 for PM; 13-2-2 for PP; and 13-3-2 for RP.

Each detector was trained using data from the given speaker. This data consists of three reference vowels: /**ɪ**/, /**ɒ**/, and /**ʊ**/.

In the testing process, speaker dependent tests and speaker independent tests were performed. Figure 7.6 reports the test result of speaker IM's *compact* feature detector tested on speaker IM. In this figure there are 6 graphs, where each graph presents the test results of 11 vowels in the context of a stop consonant. In the same way, Figures A.16, A.17 and A.18 report test results for the *compact* feature detectors for speakers PM, PP and RP, tested on these speakers themselves. Figures A.19, A.20, and A.21 are test results for the *compact* feature detectors of

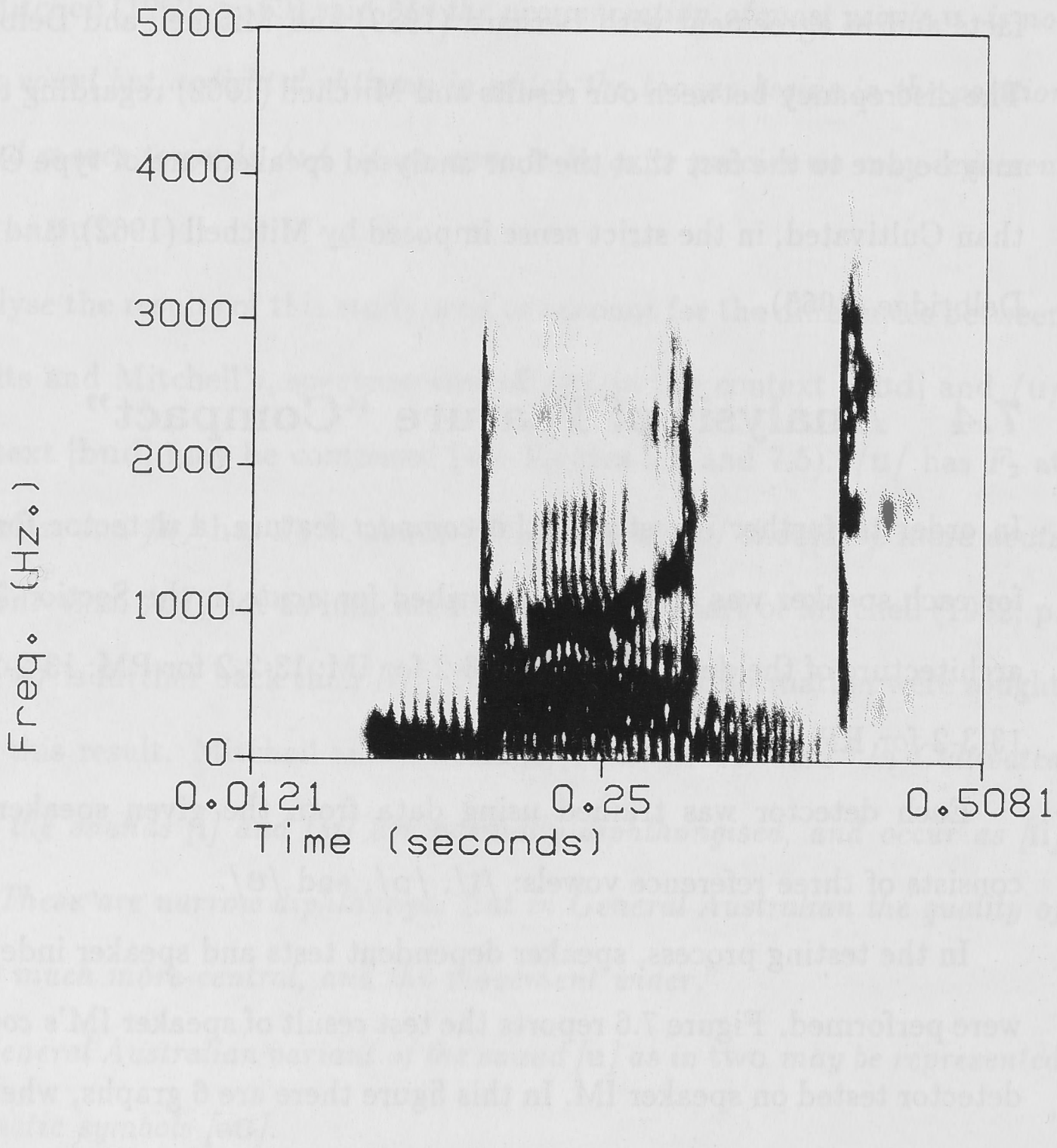


Figure 7.4: Spectrogram [**bʊd**] from speaker IM.

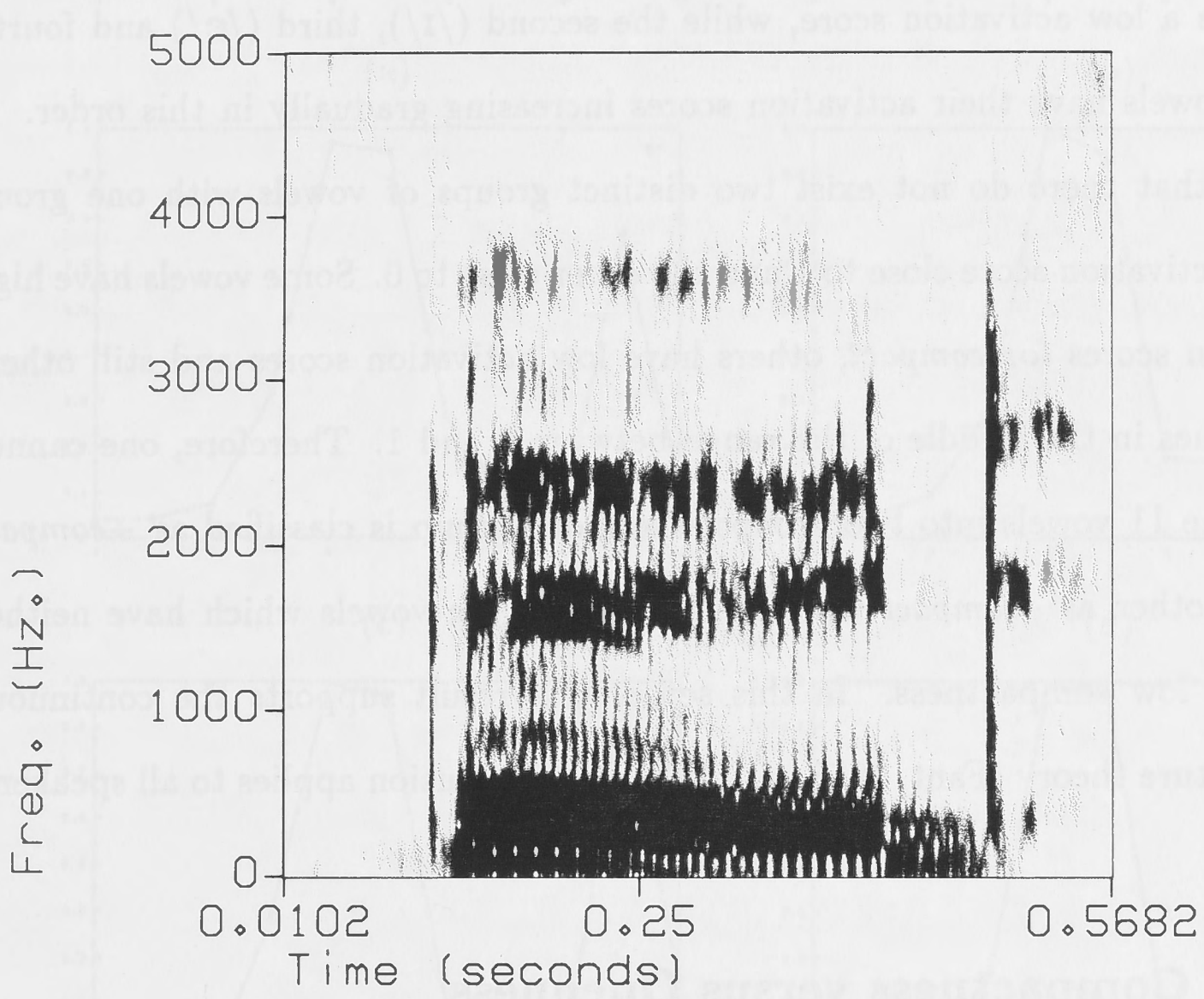


Figure 7.5: Spectrogram [**bud**] from speaker IM.

speakers IM, PP and RP tested on speaker PM. Results from the related feature *diffuse* are analysed in Appendix A.3. The following sections discuss the results of feature *compact*.

7.4.1 Binary versus Continuous

Observing Figure 7.6, this graph has two peaks representing high activation scores for *compact*, where the first peak corresponds to the fifth (/ɑ/) and sixth (/ɒ/) vowels, and the other peak corresponds to the tenth (/ʌ/) vowel. The first vowel (/i/) has a low activation score, while the second (/ɪ/), third (/ɛ/) and fourth (/æ/) vowels have their activation scores increasing gradually in this order. It is clear that there do not exist two distinct groups of vowels with one group having activation score close to 1 and the other close to 0. Some vowels have high activation scores for *compact*, others have low activation scores and still others have values in the middle of the range between 0 and 1. Therefore, one cannot divide the 11 vowels into two groups where one group is classified as *+compact* and the other as *−compact* without mistreating the vowels which have neither high nor low compactness. In this sense, this result supports the continuous value feature theory (Fant, 1973, p. 152). This conclusion applies to all speakers' results.

7.4.2 Compactness versus Openness

In articulatory phonetics, three dimensions are used to describe a vowel: (1) the height of the tongue body; (2) the front-back position of the tongue; and (3) the degree of lip rounding (Ladefoged, 1982). The height of the tongue is normally related to the openness of the mouth. The high vowels (with high tongue body) are the close vowels presented by feature *−compact* (or *+diffuse*) and the low vowels are the open vowels presented by feature *+compact* (or *−diffuse*), (see Figure 7.1 and Section 4.3).

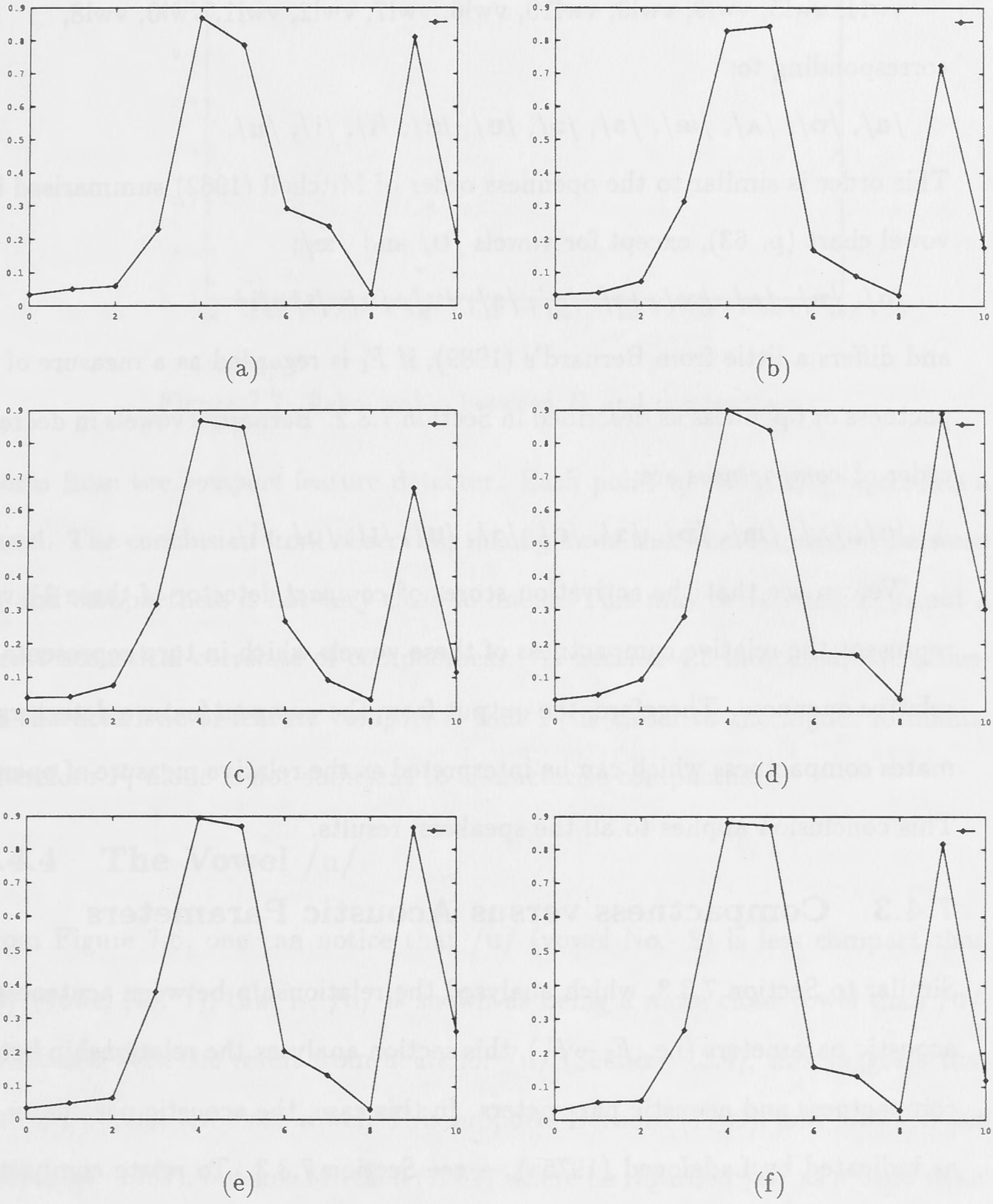


Figure 7.6: *Test results of IM's* compact *feature detector tested on IM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].

Figure 7.6 presents the relative "compactness"[3] of the vowels where the points having higher values indicate that the associated vowels are more compact. From Figure 7.6, one can place the vowels in decreasing order of compactness:

vwl4, vwl5, vwl9, vwl3, vwl10, vwl6, vwl7, vwl2, vwl1, vwl0, vwl8,

corresponding to:

/ɑ/, /ɒ/, /ʌ/, /æ/, /ɜ/, /ɔ/, /ʊ/, /ɛ/, /ɪ/, /i/, /u/.

This order is similar to the openness order of Mitchell (1962) summarised in his vowel chart (p. 63), except for vowels /ʊ/ and /ɛ/;

/ɑ/, /ɒ/, /ʌ/, /æ/, /ɔ/, /ɛ/, /ɜ/, /ʊ/, /u/, /ɪ/, /i/

and differs a little from Bernard's (1989), if F_1 is regarded as a measure of compactness or openness as described in Section 7.3.2. Bernard's vowels in decreasing order of *compactness* are:

/ɑ/, /ʌ/, /æ/, /ɒ/, /ɜ/, /ɛ/, /ɔ/, /ʊ/, /ɪ/, /u/, /i/.

We can see that the activation scores of *compact* detector of these 11 vowels represent the relative compactness of these vowels which in turn represents their relative openness. Therefore, the output from the *compact* feature detectors estimates compactness which can be interpreted as the relative measure of openness. This conclusion applies to all the speakers' results.

7.4.3 Compactness versus Acoustic Parameters

Similar to Section 7.3.3, which analysed the relationship between acuteness and acoustic parameters (i.e. $F_2 - F_1$), this section analyses the relationship between compactness and acoustic parameters. In this case, the acoustic parameter is F_1 as indicated by Ladefoged (1975a) — see Section 7.3.3. To relate compactness, which we have shown to substantially correlate with openness in articulatory terms, directly to F_1 which characterises openness (Ladefoged, 1975a, p. 173), Figure 7.7 presents a graph of F_1 versus compactness, where the horizontal axis represents F_1 and the vertical axis represents compactness on terms of activation

[3]by "compactness" I mean the MLP output after it has been trained to make "compactness" distinction.

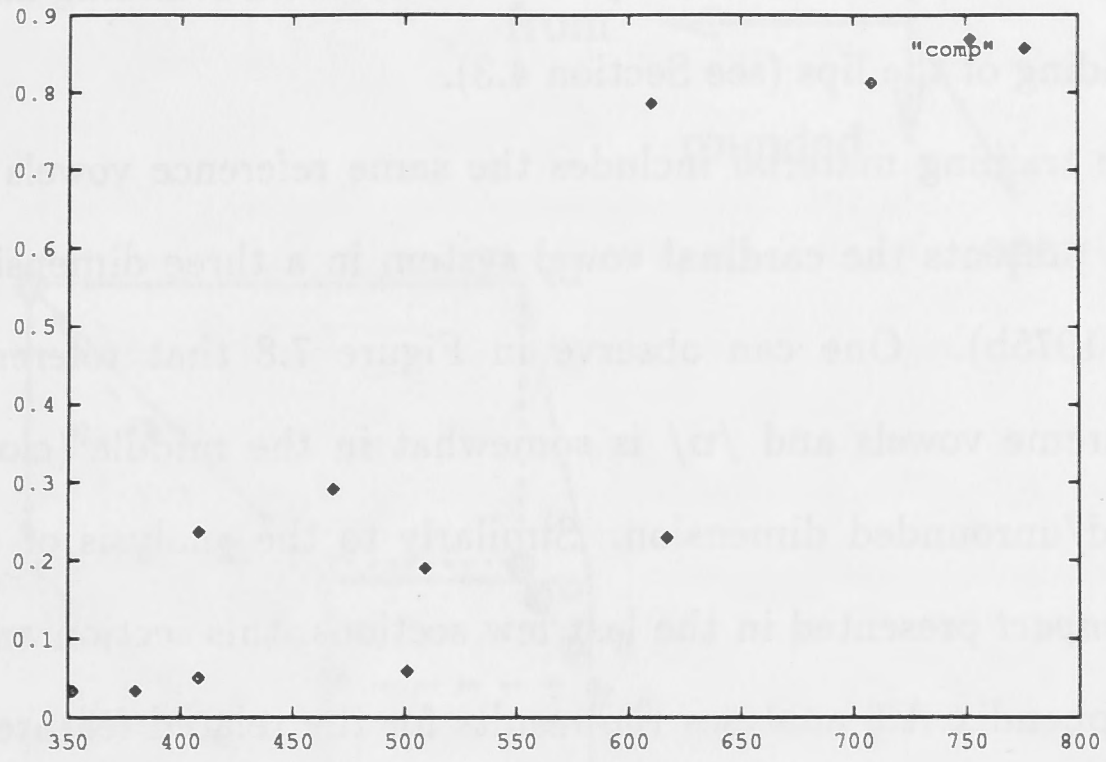


Figure 7.7: Relationship between F_1 and compactness.

scores from the *compact* feature detector. Each point in the graph represents a vowel. The conclusion from observing this figure is that the correlation between F_1 and compactness is not very close to linear. This may be because F_1 is not a direct acoustical correlate of compactness. As Section 4.3 indicated, the acoustic characteristic of feature *compact* is that F_1 is closer to the higher formants. Therefore F_1 alone is not sufficient to characterise compactness.

7.4.4 The Vowel /u/

From Figure 7.6, one can notice that /**u**/ (vowel No. 8) is less compact than /**ʊ**/ (vowel No. 7); that is, /**u**/ is shown as being a more close vowel than /**ʊ**/. Combined with the result from *acute* for /**u**/ (Section 7.3.4), this suggests that /**u**/ is a centralised close vowel. This supports Bernard (1989), and Mitchell and Delbridge (1965), but not Mitchell (1962) where he regarded /**u**/ as a back vowel in his vowel chart. This may be due to the reason described in Section 7.3.4.

7.5 Analysis of Feature "Flat"

As summarised in Section 4.3, the feature *flat* represents *a downward shifting of a set of formants or of all the formants* whereas the feature *plain* represents *no shift*

of the formant(s). The feature *flat* correlates with rounding and *plain* represents unrounding of the lips (see Section 4.3).

The training material includes the same reference vowels (/ɪ, ɒ, ʊ/). Figure 7.8 presents the cardinal vowel system in a three dimensional space (Ladefoged, 1975b). One can conclude that reference vowels /ɪ, ʊ/ are extreme vowels and /ɒ/ is somewhat in the middle (closer to /ɪ/) of the rounded/unrounded dimension. Similarly to the analysis of the features *acute* and *compact* presented in the last few sections, this section analyses the feature *flat*. Appendix A.6 analyses the results for the related feature *plain*. Figure 7.9 presents the test results of the *flat* feature detector for IM tested on speaker IM's data. Figures A.37 through A.39 present test results for the *flat* feature detector of speakers PM, PP and RP, tested on these speakers themselves. Figures A.40, A.41 and A.42 present the test results of the *flat* feature detectors of speakers IM, PP and RP tested on speaker PM. As described in Section 4.3, the feature *flat* correlates with lip rounding; the vowels that are more rounded have higher values in these graphs. The *flat* feature detector's architecture is 13-3-2 for all four speakers.

Observing Figure 7.9, each graph has one peak associated with the high activation score of *flat*. This peak is formed by the eighth vowel (/ʊ/), the next highest activation score for *flat* is the seventh (/ɔ/), and then the ninth (/u/). These last two have medial values of flatness. According to the articulatory description of the vowels presented in Section 4.4.2.1, these three vowels are all rounded vowels (the other eight remaining vowels are not rounded vowels). These results can be further checked in Figure 7.8 where the cardinal vowel system in a three dimensional space (frontness, openness and roundness; see Section 7.7) is given. We conclude that the flatness (response from the *flat* feature detectors) correlates with roundness in the articulatory space.

These results from all speakers indicate that apart from the two groups of vowels having high or low values in *flat*, there are some vowels that are neither

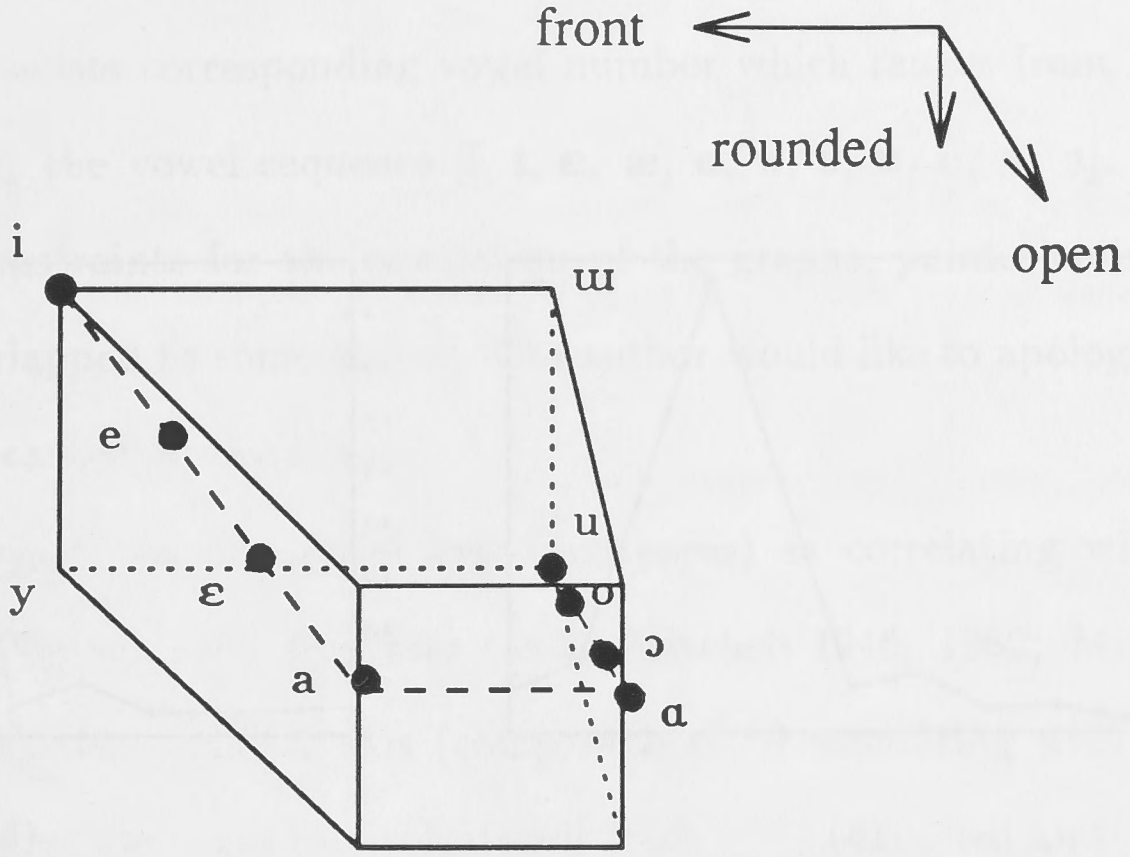


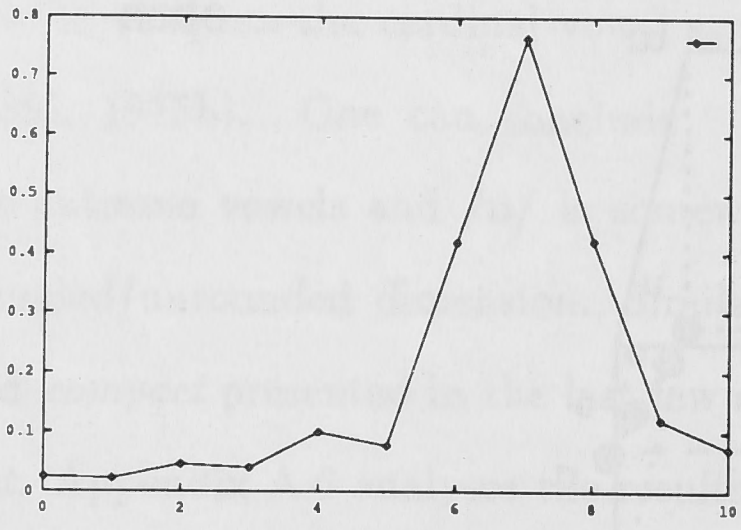
Figure 7.8: A three dimensional vowel diagram (from Ladefoged (1975b), p. 140).

high nor low, therefore, a binary decision on *flat* to categorise vowels can mistreat these vowels. These results further support the theory of the continuous nature of the features.

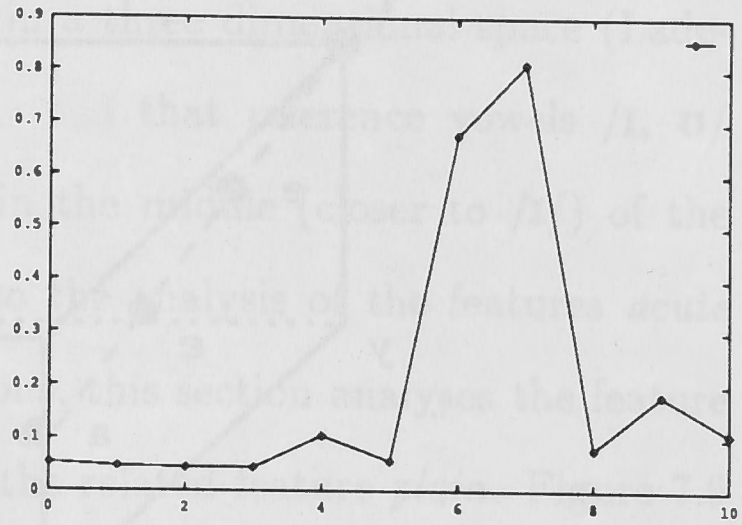
7.6 Acuteness versus Compactness: A Two Dimensional View

Sections 7.3 and 7.4 described results for the features *acute* and *compact* separately. This section presents the results of combining these two sets of results in order to compare directly with Bernard's (1989) result (in F_1 versus F_2 plane) and with Mitchell (1946, 1962) and Mitchell and Delbridge (1965) (in front/back versus open/close plane). Graphs representing acuteness versus compactness in the same plane are also included.

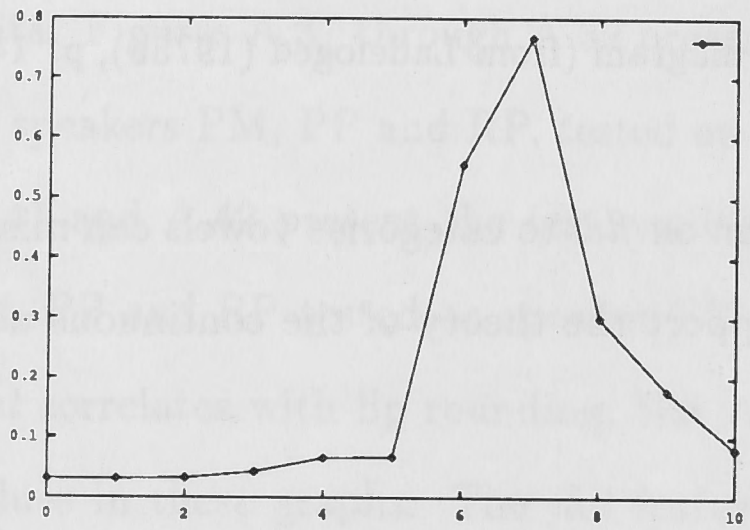
Figures 7.10 presents the test results of acuteness versus compactness, which are test results of IM's feature detectors tested on IM's data. Figures A.51 through A.53 present test results of PM, PP and RP's feature detectors tested on these speakers themselves. The horizontal axis represents the acuteness (activation score for *acute*), and the vertical axis represents compactness (the activation



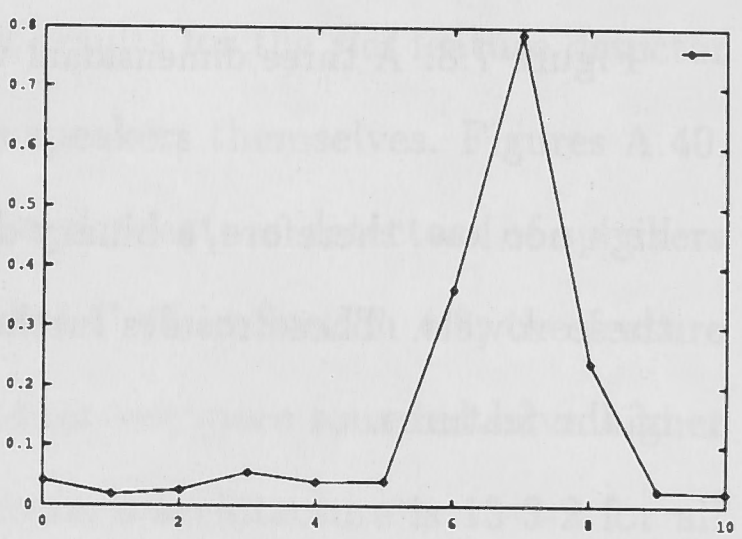
(a)



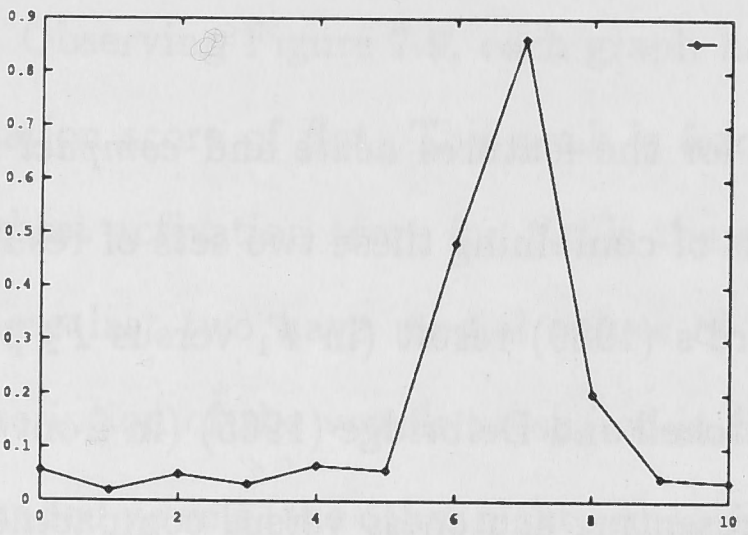
(b)



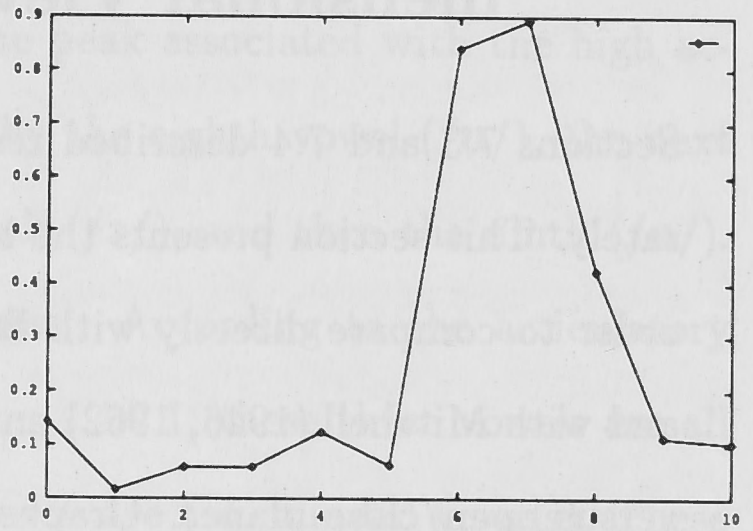
(c)



(d)



(e)



(f)

Figure 7.9: *Test results of IM's* flat *feature detector tested on IM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

score for *compact*). The symbol × represents a particular vowel and the number adjacent to it is the corresponding vowel number which ranges from 1 through 11 representing the vowel sequence [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (Because of technical constraints for the generation of the graphs, points having similar values are overlapped to some extent. The author would like to apologise for the inconvenience caused to the reader).

If we interpret the horizontal axis (acuteness) as correlating with F_2 (as in Bernard, 1989) or with *frontness* (as in Mitchell 1946, 1962; Mitchell and Delbridge, 1965), the vertical axis (compactness) as correlating with F_1 (as in Bernard, 1989) or *openness* (as in Mitchell 1946, 1962; Mitchell and Delbridge, 1965), and then move the origin of each graph from the bottom left corner to the upper right corner by turning the picture 180°, we can see a corresponding *acuteness/compactness* space, where each point represents a vowel. Observing Figure 7.10, we can see the following:

- A vowel represented in this generated feature space generally preserves its relative position in different stop consonant context, i.e. for front, back and central positions. Its absolute position may change depending on the context.

- Comparing this feature space with Mitchell and Delbridge's (Mitchell, 1962; Mitchell and Delbridge, 1965) vowel space (see the vowel chart in Figure 7.1) when we link acuteness in feature space to frontness in articulatory space, and compactness to openness, we can observe that the inter-vowel relationships (i.e. relative positions between vowels) are generally preserved, while the absolute distances between vowels may differ from Mitchell and Delbridge's. The major difference is that the vowels /ɑ/ and /ʌ/ are represented as back vowels in the feature space of Figure 7.6, while they are represented by Mitchell and Delbridge as front vowels close to central.

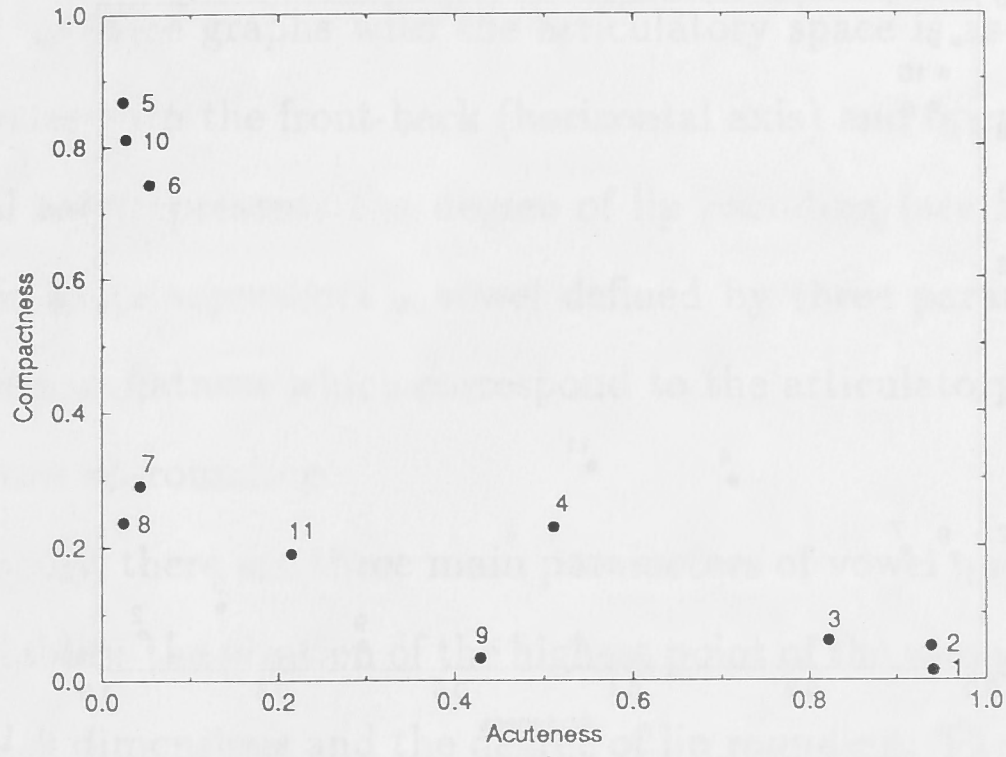
- Comparing this feature space with Bernard's (1989) F_1 vs. F_2 *space* (see Figure 7.12 on page 171), when we link F_1 to compactness, and F_2 to acuteness, we observe that the shape of the contour formed by the 11 vowels in these two spaces is different. This may be due to the fact that the relationship between this feature space and the space formed by F_1 and F_2 is that compactness correlates with F_1 and acuteness with F_2-F_1. Generally speaking, the front vowels are represented in the front for both spaces, the back vowels are located in the back and the central vowels are in the centre of both spaces. The vowel /u/ (ninth vowel) is in the centre for both cases. The vowels /ɑ/ and /ʌ/ are located far from the front vowels in Bernard's result, which correlates with the feature space to some extent. The major difference between the feature space reported here and Bernard's acoustic space is that the vowel /æ/ is located in a position which correlates with a more close and back vowel than Bernard's /æ/.

7.7 Acuteness versus Compactness versus Flatness: A Three Dimensional View

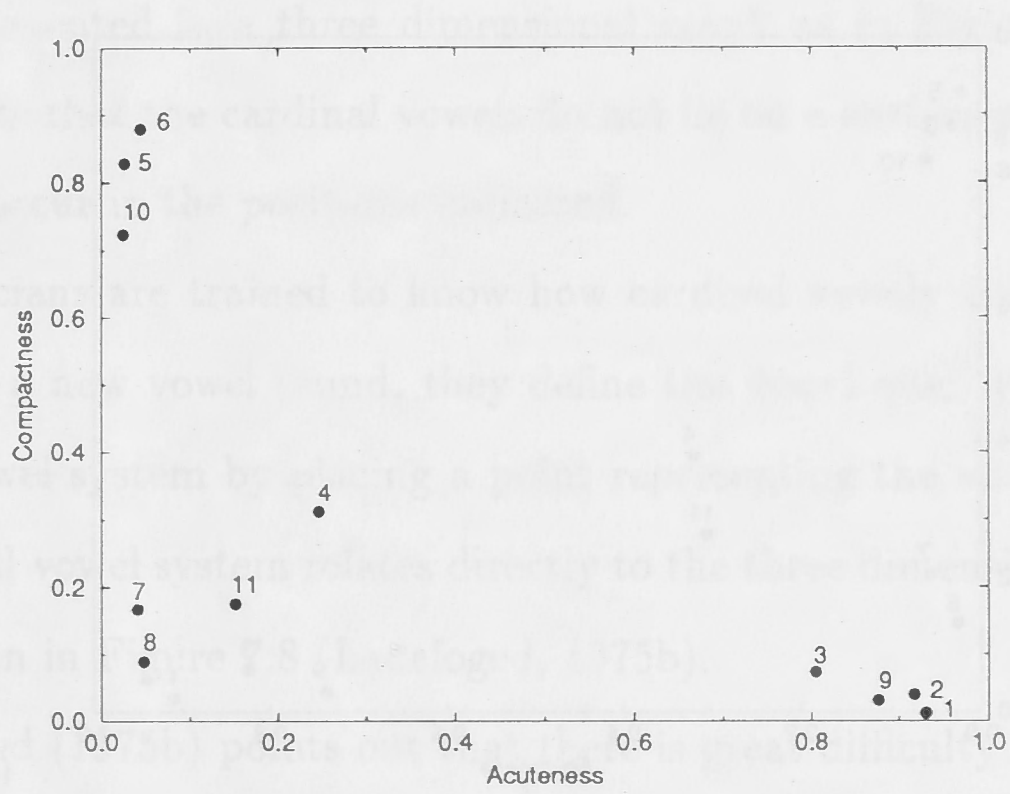
To extend the results of Section 7.6, this section analyses a combination of the responses of detectors for the features *acute, compact* and *flat*. The association between these features and the articulatory space is *acute* to front/back, *compact* to open/close, and *flat* to rounded/un-rounded (see Section 4.3).

Figure 7.11 presents the three dimensional space formed by test results on these three feature detectors of the speaker IM tested on himself. Figures A.55 through A.57 present the three dimensional space formed by test results on these three feature detectors of the speakers PM, PP and RP tested on themselves.

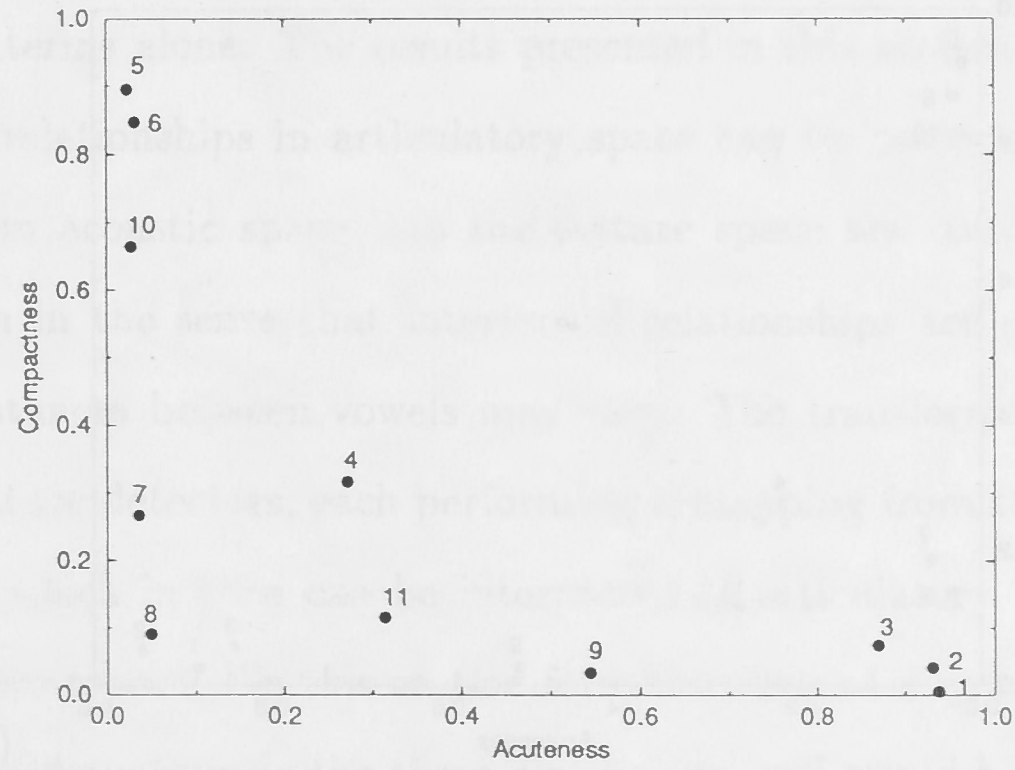
The three dimensional space is formed as follows. The base plane is formed by acuteness (activation score for feature *acute*) and compactness (activation score for *compact*) which form the horizontal axes. The vertical axis represents



(a)

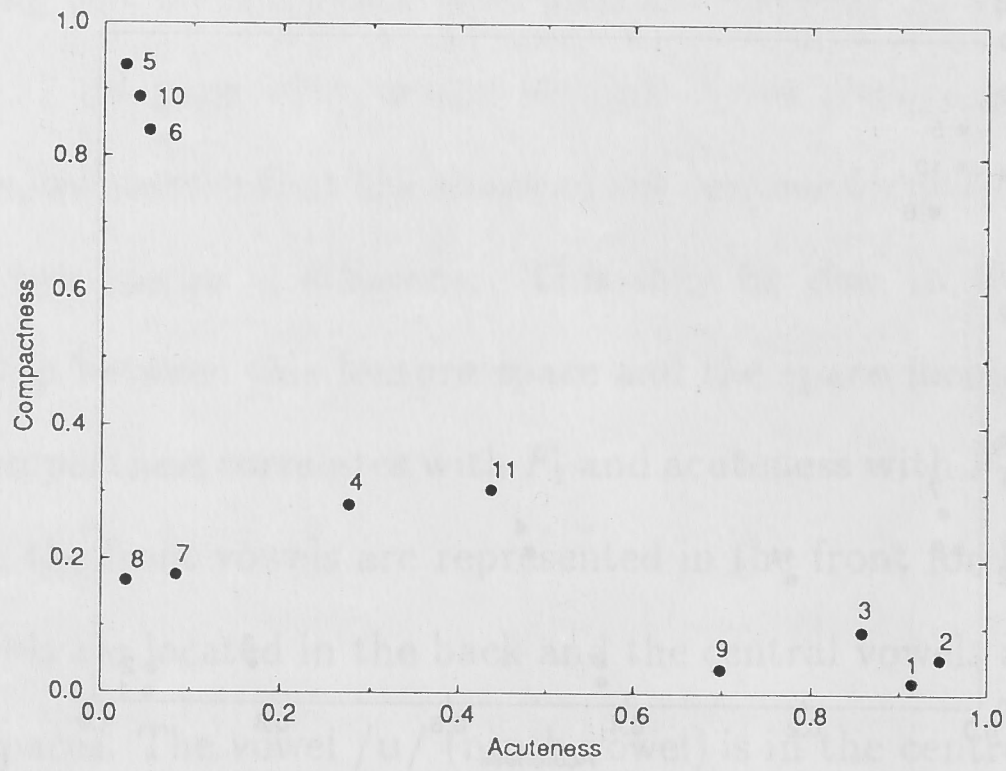


(b)

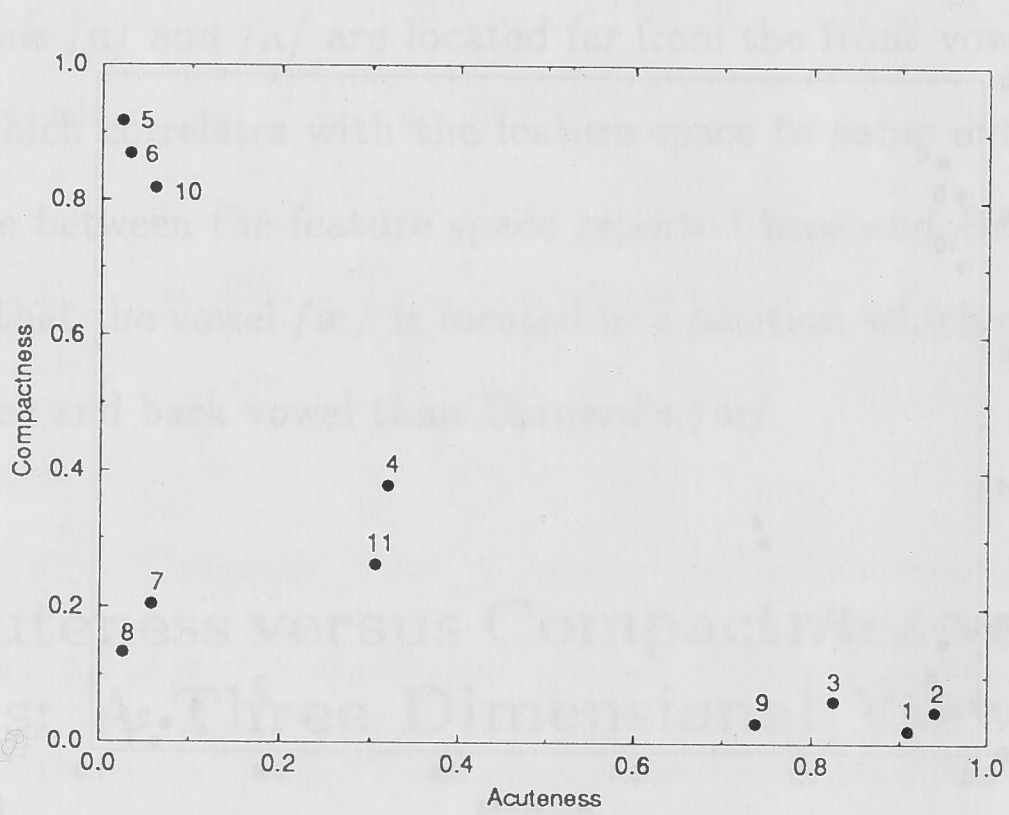


(c)

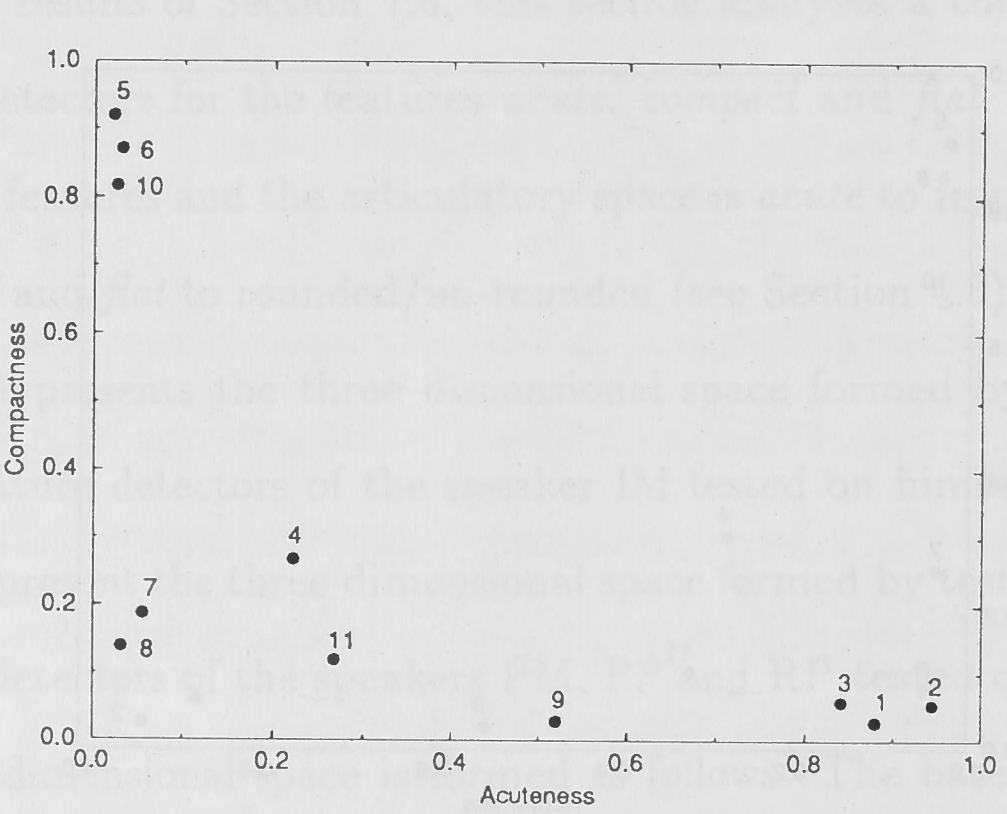
Figure 7.10: *IM's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued on page 164*)



(d)



(e)



(f)

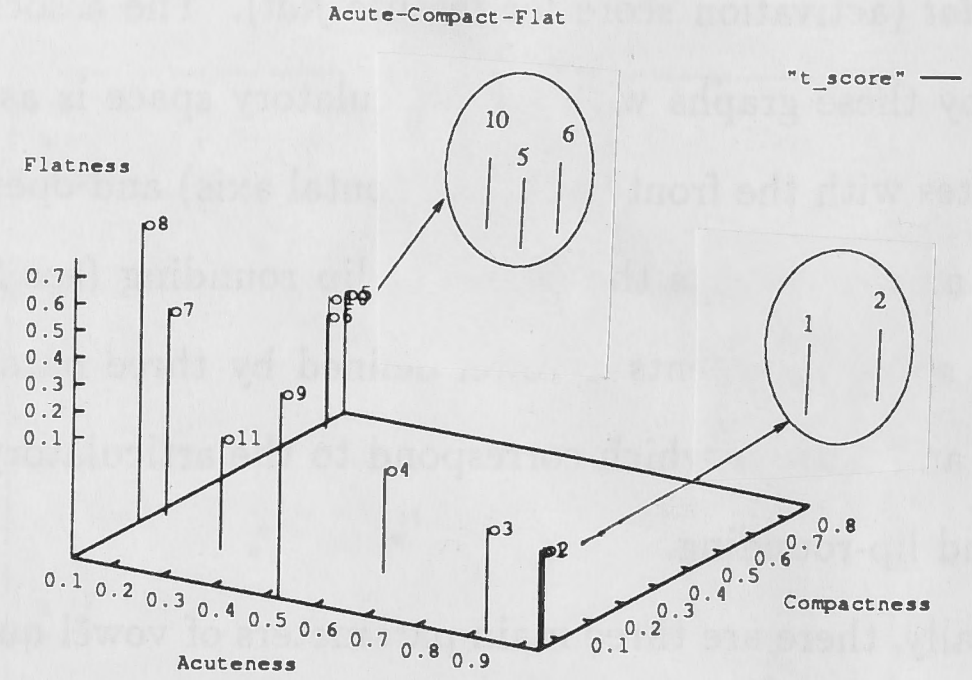
Figure 7.10: *IM's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued from page 163*)

the feature *flat* (activation score for feature *flat*). The association of the space represented by these graphs with the articulatory space is as follows. The base plane correlates with the front-back (horizontal axis) and open-close dimensions. The vertical axis represents the degree of lip rounding (see Section 4.3). Each point in this space represents a vowel defined by three parameters: acuteness, compactness and flatness which correspond to the articulatory terms front-back, open-close and lip-rounding.

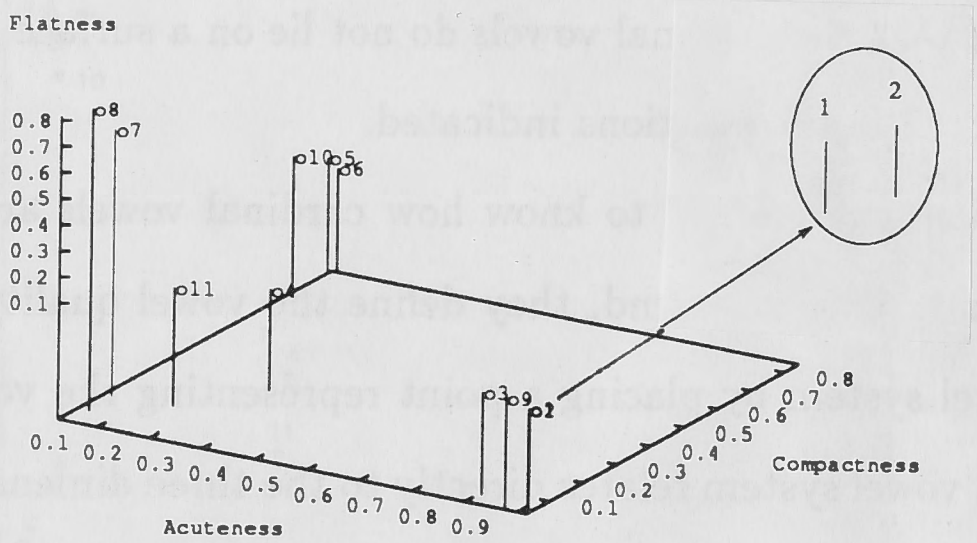
Traditionally, there are three main parameters of vowel quality that are independent variables: the position of the highest point of the tongue in the open-close and front-back dimensions and the degree of lip rounding. The quality of a vowel can be represented in a three dimensional graph as in Figure 7.8 (Ladefoged, 1975b). Note that the cardinal vowels do not lie on a surface parallel to the base plane, but occur in the positions indicated.

Phoneticians are trained to know how cardinal vowels are produced. When listening to a new vowel sound, they define the vowel quality in relation to the cardinal vowel system by placing a point representing the vowel in this system. The cardinal vowel system relates directly to the three dimensions of articulatory space as seen in Figure 7.8 (Ladefoged, 1975b).

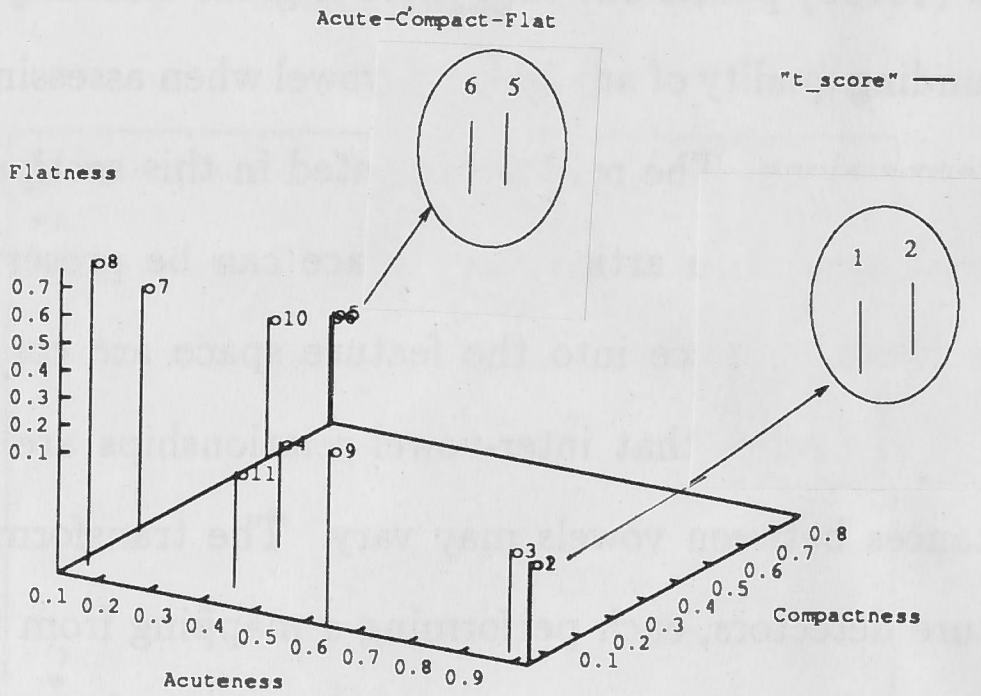
Ladefoged (1975b) points out that there is great difficulty in defining the degree of lip-rounding quality of an unknown vowel when assessing the vowel quality in auditory terms alone. The results presented in this section demonstrate that inter-vowel relationships in articulatory space can be preserved when transformations from acoustic space into the feature space are conducted, with some modification in the sense that inter-vowel relationships are preserved, but the absolute distances between vowels may vary. The transformation is performed by three feature detectors, each performing a mapping from the acoustic to feature space, which in turn can be interpreted in articulatory terms. Therefore, the techniques reported in this section form the basis of a systematic method for assessing vowel qualities in the three dimensions, and would be especially helpful in lip-rounding assessment.



(a)

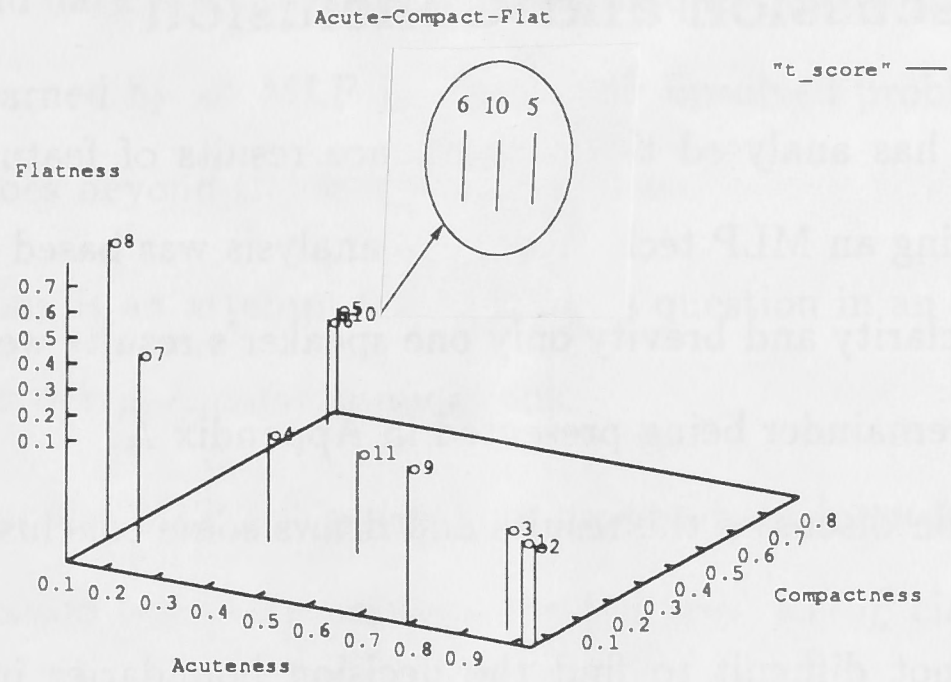


(b)

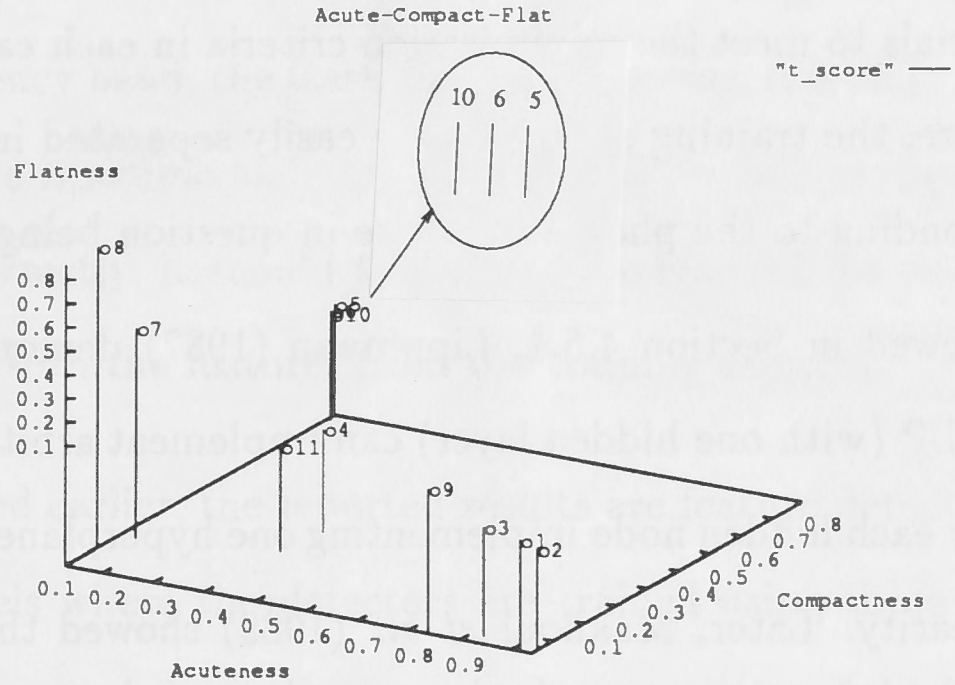


(c)

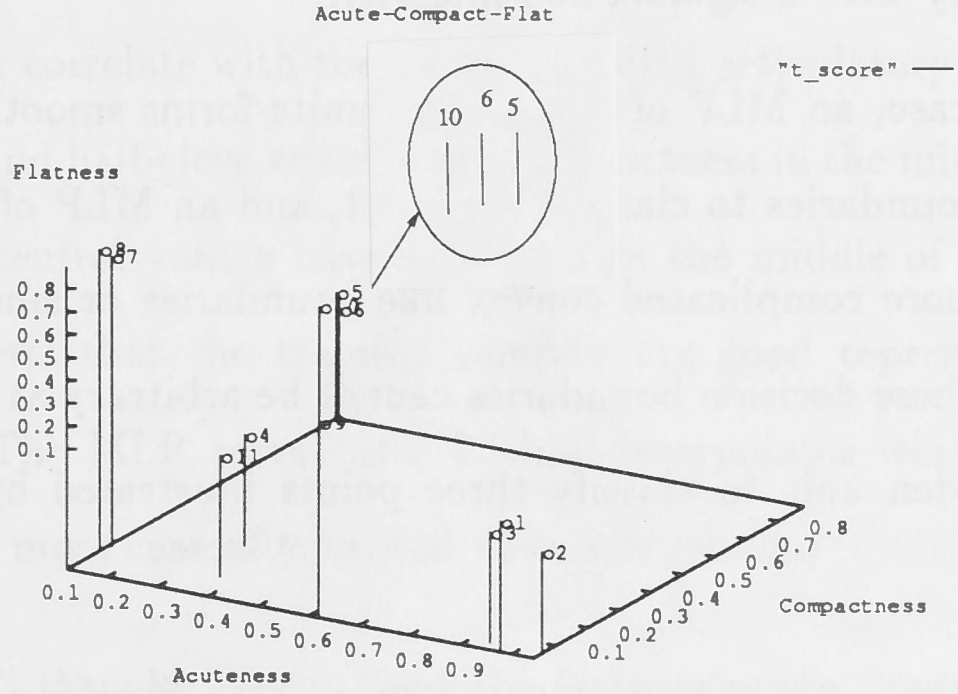
Figure 7.11: *IM's 11 pseudo steady-state vowels in an Acuteness versus Compactness versus Flatness space in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued on page 167*)



(d)



(e)



(f)

Figure 7.11: *IM's 11 pseudo steady-state vowels in an Acuteness versus Compactness versus Flatness space in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued from page 166*)

7.8 Discussion and Conclusion

This chapter has analysed the performance results of feature detectors implemented by using an MLP technique. The analysis was based on all the speakers, although for clarity and brevity only one speaker's results were presented in this chapter, the remainder being presented in Appendix A.

This section discusses the results and draws some conclusions.

- It was not difficult to find the decision boundaries implemented by the MLPs for the detectors when training, because the training did not need many trials to meet the minimisation criteria in each case (10 to 30 trials). Therefore, the training samples were easily separated into two groups that corresponding to the phonetic feature in question being *on* or *off*.

  As reviewed in Section 4.5.4, Lippmann (1987) demonstrated that a two layer MLP (with one hidden layer) can implement arbitrary convex boundaries, by each hidden node implementing one hyperplane with a hard limiter non-linearity. Later, Makhoul *et al.* (1989) showed that a two layer network can form an arbitrarily close approximation to any non-linear decision boundary with a *sigmoid* non-linearity.

  In our case, an MLP of two hidden units forms smoothed convex like decision boundaries to classify the input, and an MLP of three hidden units forms more complicated convex like boundaries or smooth closed boundaries. These decision boundaries cannot be arbitrary as in the case of using one hidden unit to classify three points illustrated by Hush and Horne (1993).

  Because very few hidden units (i.e. two or three) were needed to implement the feature detectors reported in this chapter, the MLP should have learned general properties from the training input samples instead of particular properties related to some of the samples, in which case, far more hidden

nodes would have been needed. The general question of what properties are actually learned by an MLP is a generally unsolved problem in the ANN field and goes beyond the scope of this thesis.

The following is an attempt to answer this question in an informal way, as it relates to our particular applications.

The task for the MLP in the training process is to learn from examples to draw a decision boundary between the features' *on/off* classes. By definition, there exist well defined criteria associated with the acoustic properties for each of the features (e.g. for *acute*, if the energy is concentrated in the high frequency band, the feature is on, otherwise, it is off). As the network's architecture is simple and the trained MLP is able to classify the training samples correctly, it should therefore have learned the acoustic properties associated with the features from the training samples.

- As indicated earlier, the reported results are feature detector test results of all 11 vowels where the detectors are trained using three extreme vowels. These three vowels have the acoustic characteristics of the features to a near maximum extent. The testing results of the eight vowels which are not used for training correlate with their corresponding articulatory description, e.g. half-open and half-close vowels have compactness in the middle of the range [0..1] and central vowels have acuteness in the middle of the range [0..1]. This suggests that the training samples are good representatives of the features. The MLP serves as a faithful interpolator which appropriately locates the more centrally located vowels in relation to the extreme vowels.
- Sections 7.3 through 7.5 analysed the features *acute, compact* and *flat* respectively. This analysis was based on the test results of feature detectors for all speakers tested on themselves. Similar observations can be made from cross-speaker test results. The results indicate that the vowels share

the acoustic properties associated with a particular feature to different degrees according to the position of a particular vowel in the articulatory space. (The general acoustic properties of each feature are learned by each detector from the three extreme vowels; when testing with other vowels, the detector's response represents the degree to which the testing vowels share the general properties learned by the detector.) For example, in the case of the feature *acute*, vowels towards the front are more *acute*; vowels towards the back are less *acute*. In the case of the feature *compact*, the more open vowels are more compact; the more close vowels are less compact. Although theoretically some researchers prefer a binary approach for the features (meaning that for a given vowel, a feature is present or absent), the results of this study indicate that the binary approach supported by some researchers (see Section 3.2.6) mistreats vowels whose feature values are neither high, nor low. This suggests that the features have more than two values and tend to be continuous in nature.

The continuous nature of different features is associated with different dimensions of the articulatory space. Specifically, the *acute* feature correlates with the front/back dimension, the *compact* feature associates with the open/close dimension and the *flat* feature associates with the rounded/unrounded dimension. The relationship between vowels in the feature space is similar to their relationship in the articulatory space. This point was further illustrated in Sections 7.6 and 7.7 where a two dimensional feature space and a three dimensional space were established explicitly.

In view of the strong correlation between the feature and articulatory spaces, the techniques examined in this chapter provide a mechanism for mapping from the feature space to the articulatory space. These techniques provide a systematic way of categorising vowels in articulatory/feature terms from the acoustic parameters, which is very helpful for people who are not trained phoneticians and for training of the new phoneticians. These techniques are

specially helpful when assessing the degree of lip-rounding of vowel quality, because of the recognised difficulty of this task. Before such techniques are put into the practice, they can be evaluated by comparing the results obtained with judgements of the well well-trained phoneticians, given the same series of sounds.

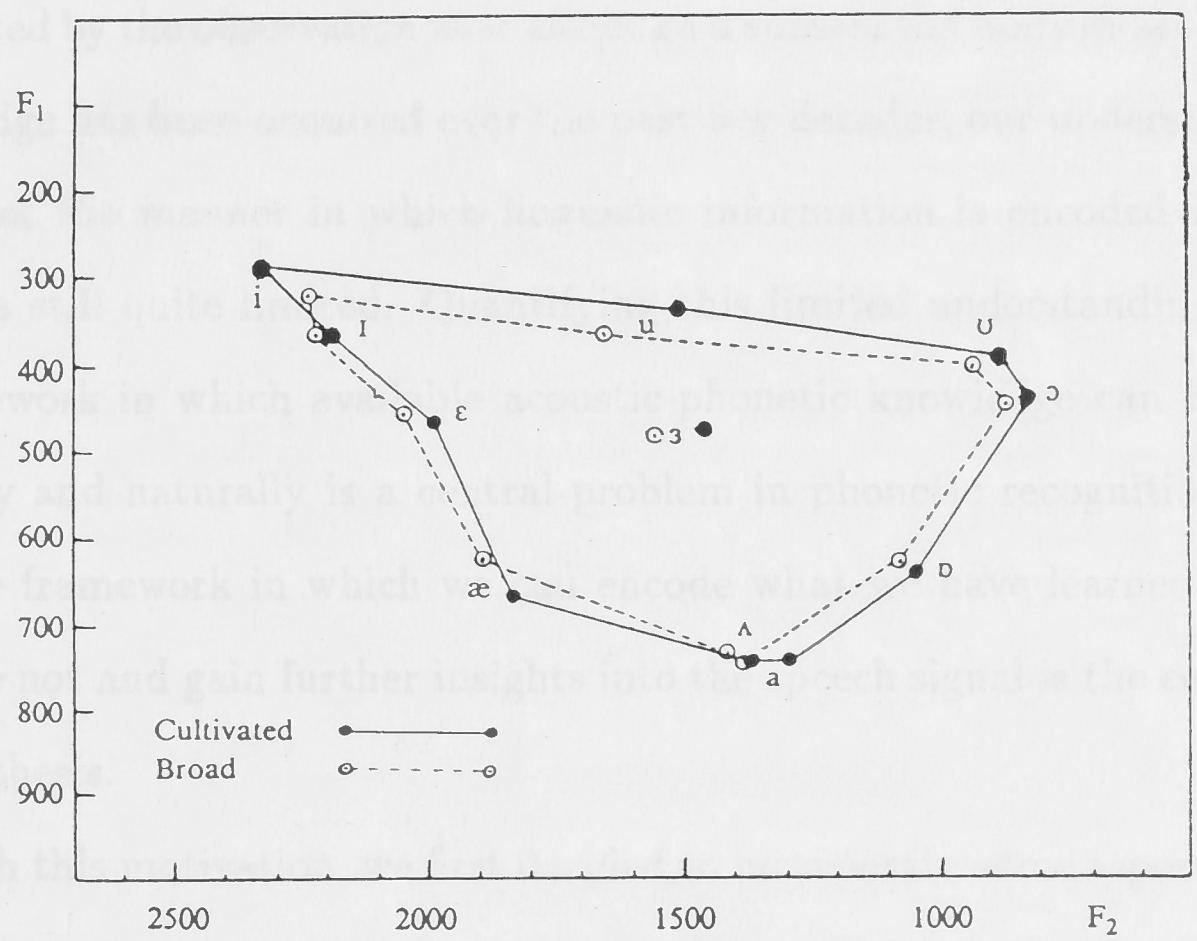
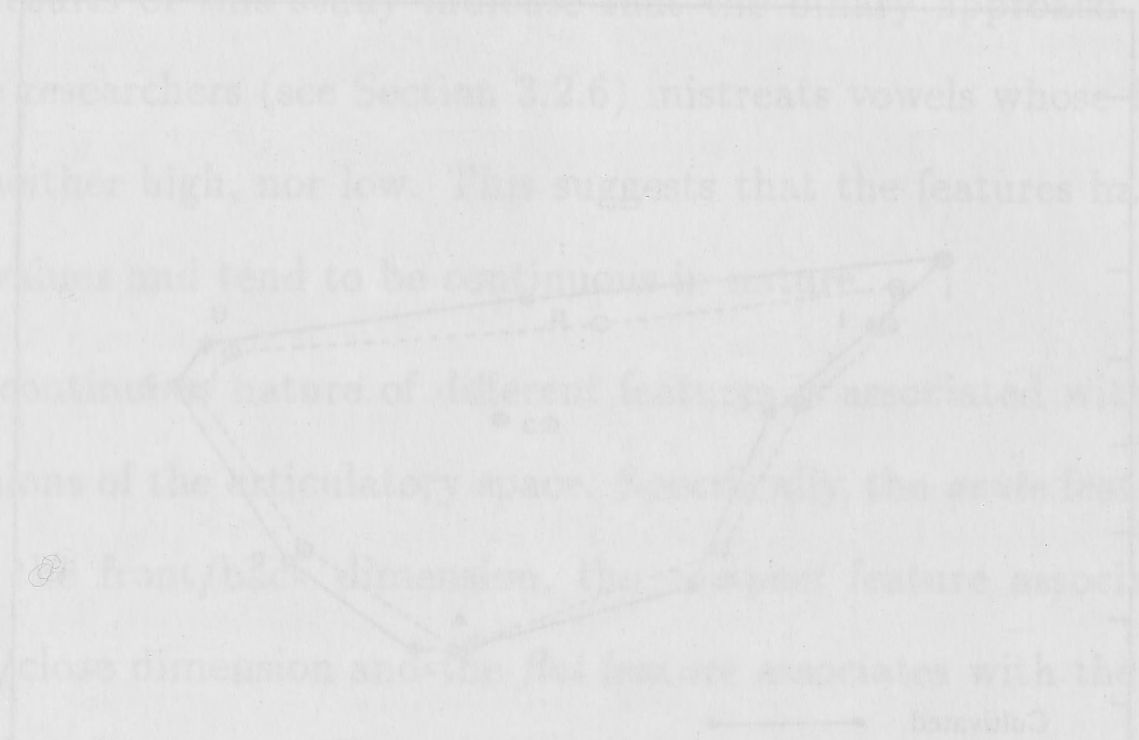


Figure 7.12: Bernard's (1989) F_1 vs. F_2 space (page 194).



Chapter 8

Discussions and Conclusions

8.1 Summary

This thesis presents an investigation into the encoding of speech knowledge for speech recognition using modularly organised artificial neural networks (ANNs), motivated by the observation that although a substantial body of acoustic-phonetic knowledge has been acquired over the past few decades, our understanding of the details of the manner in which linguistic information is encoded in the speech signal is still quite limited. Quantifying this limited understanding and finding a framework in which available acoustic-phonetic knowledge can be utilised effectively and naturally is a central problem in phonetic recognition. Finding a suitable framework in which we can encode what we have learned, model what we have not and gain further insights into the speech signal is the central concern of this thesis.

With this motivation, we first decided to incorporate some aspects of phonetic knowledge into the process of speech recognition. The specific phonetic knowledge chosen here is the *distinctive features* of Jakobson, Fant, and Halle.

The speech material used for this study was $<stop><vowel>[d]$ words where $<stop>$ and $<vowel>$ were varied. In order to relate $<stop>$ and $<vowel>$ to Jakobson *et al.*'s features in an Australian English context, we established feature based definitions for these phonemes.

The following sections summarise our explorations and discuss some issues that have been raised.

8.2 Phonetic Recognition

The given set of distinctive features was incorporated into speech recognition via a three level modular system. The first level detected the features *voicebar* and *vocalic*, and divided the speech signal into two major classes: *vocalic* and *non-vocalic*. The second level detected fine features within these major groups; each feature was associated with one feature detector and detectors worked in parallel. The third level performed phonetic classification on the basis of the feature detection results, using two modules which handled the *vocalic/non-vocalic* classes separately.

In order to select the best approach to feature detection, we compared two schemes: in one scheme, burst-onset intervals and pseudo steady-state vowels shared the same feature detectors; in the other scheme, these two types of speech material had their own feature detectors. The results of this comparison showed that the second scheme was better. The implication is that acoustic realisations of the same features differ in these two types of speech material.

In order to evaluate different techniques for phonetic recognition after feature detection at the third level, we experimented with Multi-Layer Perceptron (MLP) and table look-up techniques and found that the MLP technique which performed phoneme recognition using continuous features gave better results than the table look-up technique which performed phoneme recognition on the basis of binary features. This result provides evidence in support of the view that the features have a of continuous rather than binary nature.

We also compared phonetic recognition using an explicit distinctive feature detection approach with two monolithic approaches — one performed recognition of 17 phonemes directly from the input acoustic parameters and the other pre-classified the input into *vocalic* and *non-vocalic* classes and the phoneme recognition was then performed within each class. The approach proposed in this thesis (using explicit feature detection prior to recognition) and an approach which pre-classified the phonemes into *vocalic* and *non-vocalic* groups both performed

better than an approach which used a single monolithic module. The proposed approach performed better than the approach with two monolithic modules for vowel recognition, and had similar results for stop consonant place distinction. The results from this study showed that one advantage of having explicit feature detection prior to phoneme recognition is that the performance can be analysed by relating it to the system design at an individual module level. This facilitates diagnosis and correction of problems. A good example was found while analysing the recognition results. We found that the main errors occurring in recognition were due to the poor performance of the *tense/lax* feature detectors. This in turn was a consequence of the distinctive features' static framework. Elaboration of this framework to take account of speech dynamics will improve these and other feature detectors' performance and overall recognition results. The *tense/lax* detectors can then be modified independently of other feature detectors modules, with a potential halving of recognition errors.

The feature detection results demonstrate that it is feasible to detect a representative set of distinctive features from the acoustic signal on a speaker independent basis. Although these results are based on limited data, they give broad support to the theory of invariant cues being present in the acoustic signal.

This study was based on speech material from four speakers. Although we expect the approach investigated in this thesis to be a speaker independent one, more speakers need to be tested to establish this more thoroughly.

8.2.1 Feature Tense/Lax

As indicated in Chapter 6, the analysis of the recognition performance shows that many of the errors which occurred were due to the *tense/lax* detectors' low detection rate. The low detection rate of this feature pair is due to the limitation of the distinctive feature's static framework, where only static information is used to characterise each feature. Lifting this limitation and utilising dynamic information relevant to this pair of features should improve their detection rate

significantly, yielding significant improvement in the recognition rates for vowels and stop consonants. The error rate for recognition of vowels and stops can be reduced up to 50%. The dynamic information component of these features is related to the duration of the vowels and the voice onset time for the stop consonants. Possible ways of lifting the static constraint of the distinctive features framework and integrating such timing information into these feature detector models are (i) for the vowels, to measure the period of time over which the current fine feature patterns is active; (ii) for the stops consonants, incorporate the voice onset time into the model.

8.2.2 Taking Explicit Account of Speech Dynamics

This study has concentrated on recognition using the steady-state portion of vowels and the burst onset-interval of the stop consonants. Test results of recognition during the transitional period using the *vocalic* and *non-vocalic* models derived from steady-state vowels and burst-onset intervals show that the signal near to the burst-onset interval has more stop consonant information than the signal which is far from it, This finding is quite expected from the theoretical point of view. How to best model the transitional information taking this information into account still needs further investigation. Stevens (1986) indicated that this contextual information could be represented in terms of distinctive features which could potentially allow the contextual variability to be accounted for in a more natural manner. Stevens (1992) further suggested that lexical items may be represented in terms of the features. How to put this theory into practice, that is how to define the transitional portion of signal with distinctive features and how best to extract the features from steady-state and transitional portion of signal still needs further research effort.

8.2.3 Extending Phonetic Scope

This study has been based on relatively limited speech material: $<stop><vowel>[d]$. An immediate question arises: how could this study be extended to give a wider phonetic coverage, that is to include more speech sounds.

The present study modelled a set of feature detectors which are relevant to the speech material included. To extend the speech material, other features which are relevant in categorising the new phonetic material have to be added. Therefore, new detectors for these added features have to be created.

In order to implement these new detectors correctly, one has to analyse what the acoustic characteristics of each of the new set of features are, and decide whether to add further types of acoustic representations. The present study uses Linear Predictive Cepstral Coefficients (LPCCs) as the input acoustic representation, because all the features included are based on spectral information (except that part of *tense/lax* which depends on duration or voice onset time). LPCCs may not be adequate for characterising the new features. For example, if the extended set of phonetic information includes nasal consonants, then the feature *nasal/oral* has to be added to distinguish nasals from orals. Because of the presence of zeros in the spectrum, a low order LPCC cannot represent nasals accurately and other forms of acoustic representation will have to be considered.

8.2.4 Continuous Speech

Another immediate question is how to extend this study to cope with the difficult case of continuous speech. The present study has concentrated on burst-onset intervals and the steady-state portion of vowels in CVC syllables. In order to extend this study to continuous speech, the first problem to be solved is extension to a broader phonetic coverage, as addressed in Section 8.2.3. On the basis of an extended system covering the necessary phonetic context, *vocalic/non-vocalic* detectors could serve as on-line segmenters which classify the incoming speech into *vocalic/non-vocalic* groups. The fine feature detectors of each group could detect

the fine features in a similar manner to that described in earlier chapters. The second question which needs to be addressed is that of transitional information processing, as addressed in Section 8.2.2. The third question is that the proposed approach needs good segmentation and labelling for the training material. This problem could be approached by employing other techniques such as HMM.

8.3 Feature Analysis

In order to investigate some general properties of features, we examined the feature detector responses when the detectors had been trained using three reference vowels. We analysed the performance of the detectors on all 11 vowels.

8.3.1 Binary versus Continuous

One property of concern regarding distinctive features was whether they are *binary* or *continuous* in nature. We addressed this question by a mapping of acoustic information of eight vowels not used in the training onto a feature space generated by three reference vowels. The results indicate that the vowels share the acoustic properties associated with each feature to different degrees; there was no obvious division into two classes (e.g. *present* vs. *absent*) which would be the outcome with binary features. This in turn indicates that the features are continuous in nature, although they have been treated as binary by many researchers. This indication was further strengthened by recognition experiments in which an MLP based approach after feature detection was compared with a table look-up approach. The MLP based approach gave a much better result than the table look-up, the major difference between these two approaches being that the table look-up was based on a *presence/absence* binary decision of the features, whereas the MLP approach was based on degrees to which features were present (ranging from 0 to 1). Thus, both the mapping of vowels onto the feature space and the efficacy of continuous features in vowel recognition point to the continuous nature of the features.

8.3.2 Feature – Articulatory Correlation

This study also analysed the correlation between distinctive features and the articulatory description of vowels. The results showed that each feature pair correlates well with one dimension of articulatory space. Specifically, *acute/grave* correlates with the *front/back* dimension, e.g. the front-most vowels have higher *acute* confidence values, the back-most vowels have lower *acute* confidence values. The *compact/diffuse* correlate with the *open/close* dimension. The more open vowels have higher confidence values in *compact* and the more close vowels have lower confidence values in *compact*. The *flat/plain* distinction correlates with the *rounded/unrounded* dimension. The more rounded vowels have higher *flat* confidence scores, and the less rounded vowels have lower *flat* confidence scores. This aggregated results support Jakobson *et al.*'s distinctive feature definition in articulatory terms.

8.3.3 Acoustic Space to Articulatory Space Mapping

On the basis of the results of feature analysis, a feature space represented by the features *acute, compact, flat* was established. In this feature space, vowels are represented in three dimensions. The conclusion from observing the 11 vowels studied in this feature space is that inter-vowel relationships generally correspond to inter-vowel relationships in an articulatory space, although exact inter-vowel distances may not be preserved. These results further demonstrate that the feature detectors provide a mapping from an acoustic space to an articulatory space, helping to determine relative relationships between vowels, e.g. to find which of two vowels is more fronted, or more open, or more rounded. This development provides an objective method for assessing vowel quality. Further studies may acquire four cardinal vowels (lying towards each corner of the cardinal vowel chart) as reference vowels instead of using three extreme vowels as references. This should result in a more accurate mapping from acoustic to articulatory space, where not only inter-vowel relationships in articulatory space should be revealed,

but a more accurate location of each vowel should also be obtained. If this prediction is confirmed, then this technique will provide a very good tool for assessing vowel quality, especially for people who are not trained phonetically and also in training new phoneticians.

8.4 Knowledge-Based and Pattern Recognition: A Combined Approach

Ongoing debate in the literature shows that our understanding of the speech code is rather limited. It is more difficult still to quantify this knowledge in many cases. The approach adopted herein combines knowledge-based and pattern recognition techniques where we model the knowledge (distinctive features) with the system structure, and use pattern recognition techniques (artificial neural networks) to cope with our inability to directly quantify the defining properties of individual features. The proposed modular architecture provides a suitable framework to encode these forms of speech knowledge which we can make explicit. The use of ANNs helps us cope with our lack of quantitative knowledge regarding acoustic properties of the distinctive features. In this modularly organised MLP framework, the feature detectors can be trained and used in parallel. The intermediate results in the recognition process can be analysed easily, facilitating the location and rectification of any problems. Independent improvements can be made to inadequate modules without the need to modify other modules. This process has been demonstrated in the case of the feature pair *tense/lax*.

8.4.1 Multi-Layer Perceptrons

The Multi-Layer Perceptron (MLP) type of ANN is the basic pattern classification technique used throughout the present study. The results demonstrate that MLPs provide a very flexible framework for integrating speech knowledge into an ASR system. They need however, to be modularised in order to reduce the complexity of each module. We use relevant speech knowledge to determine an overall

hierarchical structure of MLP modules. It is an useful technique to make use of our knowledge, and to model what we do not know. It also helps us to discover more insight into the speech code. Used appropriately, MLPs have proved to be very useful for feature detection where we do not have quantitative knowledge of the features. They extract general relationships from the training samples. By using them, we performed the recognition task, the analysis of distinctive feature properties and analysis of the distinctive feature space to articulatory space relationship. We expect to further discover what has been learned by the MLP, in order for them to detect the features. This will lead us to a better understanding of the speech code.

8.4.2 Initial Conditions

Our experience in using MLPs is that they can be very sensitive to the initial conditions. Different initial conditions may lead to different local minima on the error surface, some of which may not be good enough to solve a given problem. We followed Lippmann's (1987) suggestion, by choosing different small random initial conditions and selecting the trained MLP with the best performance. The overall performance of the trained MLPs did vary noticeably with different initial conditions. Although Lippmann's suggestion is a practical solution at present, it is expensive computationally and a better theoretical understanding of the problems encountered here is desirable.

8.4.3 Architecture Selection

The question of what is the most suitable practical architecture for a given problem is still unanswered (Hush and Horne, 1993). Our approach concurred with Hush and Horne's suggestions here, starting with a very simple architecture and increasing its size if the result was unsatisfactory. This process was repeated until a satisfactory solution was found.

Two other techniques could have been used, and need further exploration. One of these techniques is *cascade correlation* (Fahlman and Lebiere, 1990). The

idea is to start with one node and create additional nodes as they are needed. These additional nodes are created during training.

The other technique is to start with a large network and then apply a *pruning technique* which eliminates weights and/or nodes which contribute little or nothing to the solution (Le Cun *et al.*, 1990). One must however have some idea of what size of network constitutes a "large" network. It is obvious that theoretical investigation of the architecture selection for a given problem is required.

8.4.4 Internal Representations

Because we lack an adequate quantitative understanding of the acoustic properties associated with each feature, we employed MLPs to learn from training samples. The properties learned by the MLPs to perform feature detection have not as yet been determined. Further studies should investigate this by looking into the internal structure of the MLP (e.g. weights and functions), to discover in what ways the internal representations of the MLP relate to possible acoustic properties. If successful, this will help us to gain some quantitative views of the acoustic properties of the features, and which in turn will help us to gain further insights into the speech code.

8.5 Conclusions

Based on this thesis study, we conclude the following.

- By combining knowledge-based and pattern recognition techniques, we can make use of what we have learned, model what we have not and further gain new insights into the nature and details of the speech code.

- By modelling distinctive features in a modularly organised system, the modules can be trained and used in parallel. We can analyse the system performance in terms of that of individual modules, making the system performance more understandable. Intermediate results in the recognition

process can be analysed easily, helping to locate and solve any problem. Independent improvements can then be made to inadequate models without the need of modifying other modules. This approach also reduces the system complexity by using a modular organisation. We give MLPs "bite-sized" solvable problems which can be resolved by using simple architectures trained by using small data set. We expect that this will facilitate our understanding of the properties learned by the MLPs in the future.

- We have demonstrated that distinctive features can be extracted from the acoustic signal which provided evidence in support of the invariance existence theory. Explicit distinctive feature detection prior to recognition aids in the recognition of vowels and stop consonants.

- Jakobson *et al.*'s distinctive feature's static framework constrains recognition performance. This framework needs to be elaborated to allow further improvement of recognition results.

- We demonstrated that the realisation of the distinctive features in the speech signal is continuous in nature rather than binary.

- We also demonstrated that Jakobson *et al.*'s distinctive features are directly related to articulation for vowels. A mapping from the distinctive feature space, generated by three reference vowels, to the articulatory space was established by using feature detectors to implement the mapping function. The results of this mapping are in accord with inter-vowel relationships in the articulatory space. With further verification using four extreme cardinal vowels, this technique has the potential to provide objective assessment of the articulatory quality of vowels.

Appendix A

Other Related Issues of Feature Detectors' Analysis

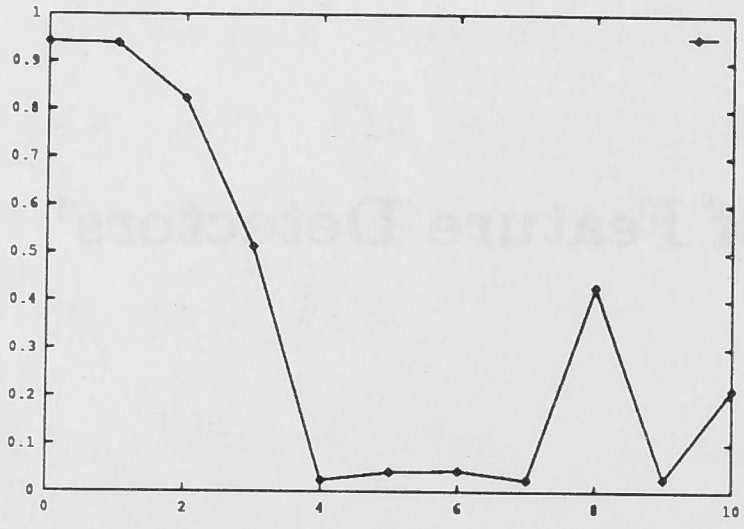
For clarity and conciseness of Chapter 7, we divided the material to form Chapter 7 and this appendix. As indicated in Section 7.1, we present the results and analysis of features *acute, compact* and *flat* in Chapter 7 with figures from one speaker only. This appendix consists of a complete set of feature detector test results from all the speakers described in Chapter 7, including the results for complementary features (i.e. *grave, diffuse* and *plain*). Therefore, the set of results presented here are test results for detectors of features *acute, compact, diffuse, grave, flat*, and *plain* for speakers IM, PM, PP and RP, tested on these speakers themselves and on PM respectively (see also Chapter 7). This appendix also makes further observations which do not directly contribute to the main points made in Chapter 7.

A.1 Feature "Acute"

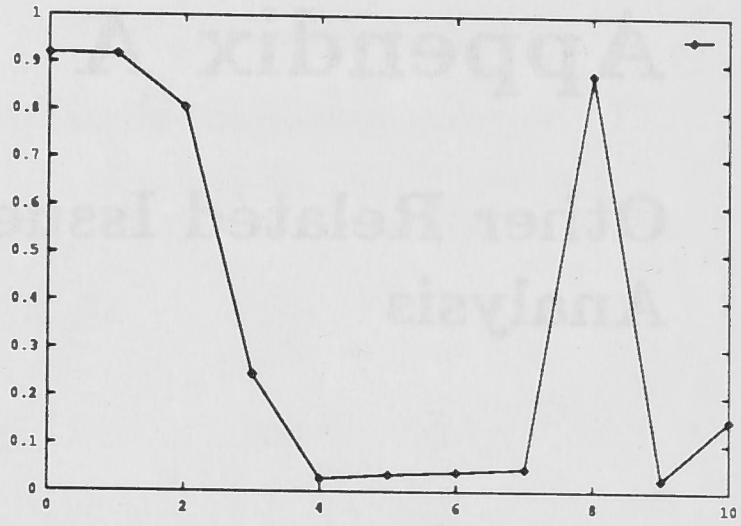
Figures A.1 through A.7 present a complete set of test results of feature detector for *acute*. Some additional observations can be made as following:

- The observations made in Section 7.3 relating to speaker dependent tests also apply to the speaker independent test results.

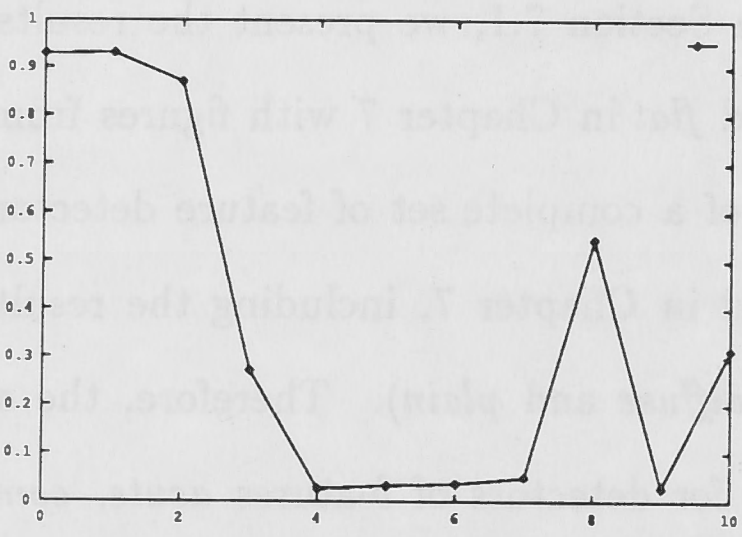
- From Figure A.1, one can clearly observe that /ɑ/ (vowel No. 4) and /ʌ/ (vowel No. 9) both have lower activation scores than /ɜ/ (vowel No. 10),



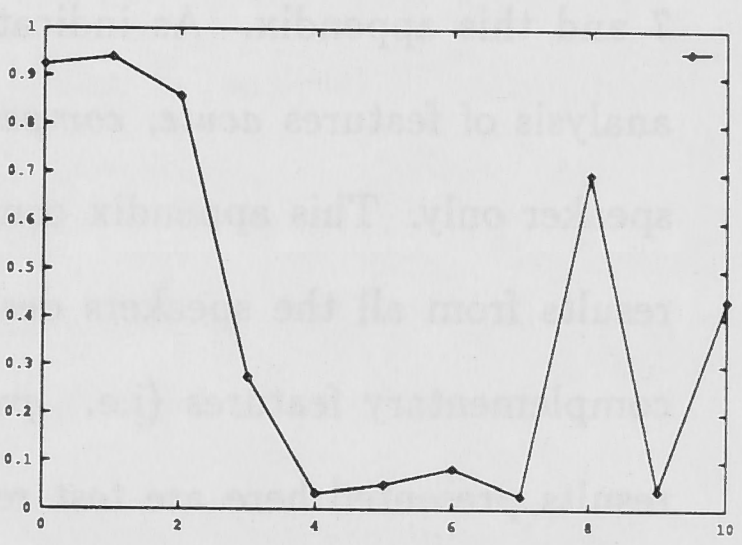
(a)



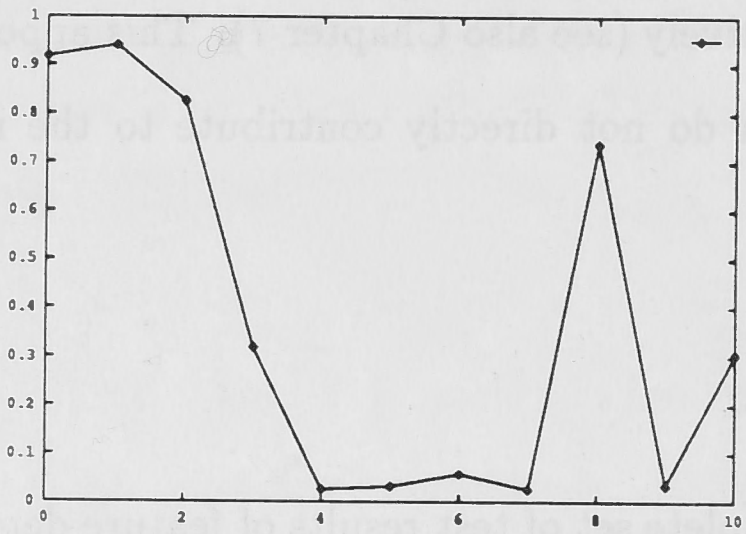
(b)



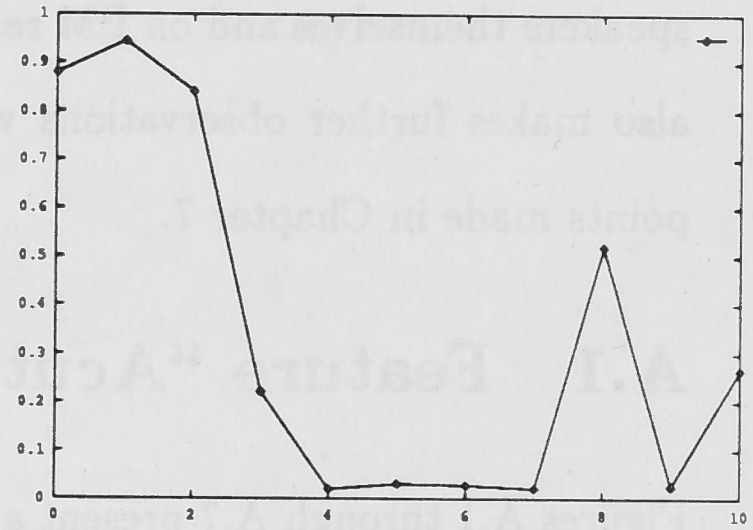
(c)



(d)



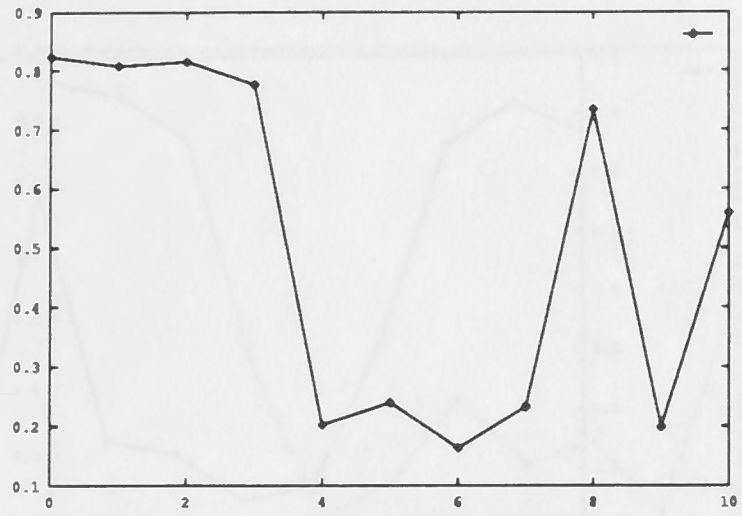
(e)



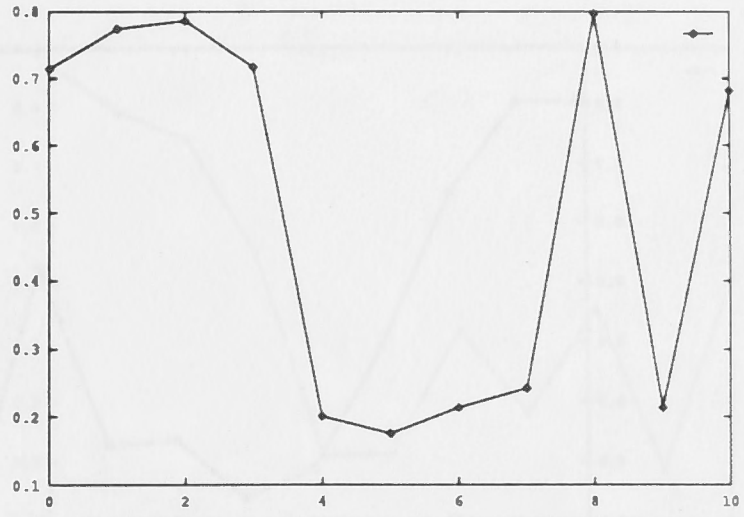
(f)

Figure A.1: *Test results of IM's* acute *feature detector tested on IM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]

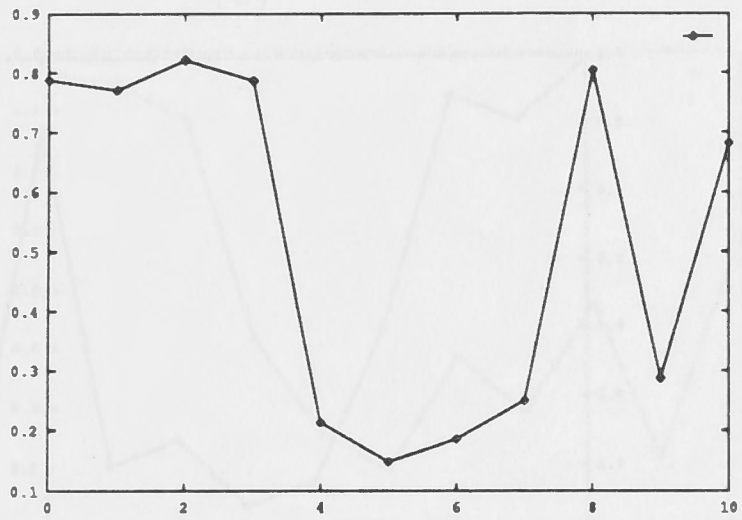
.



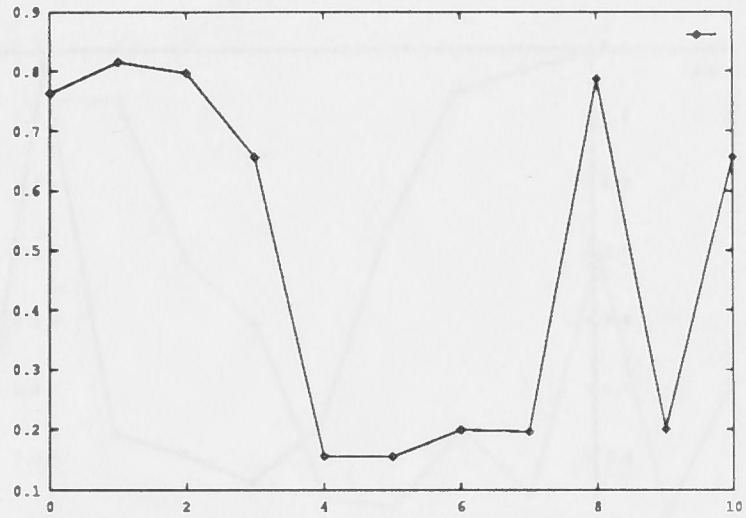
(a)



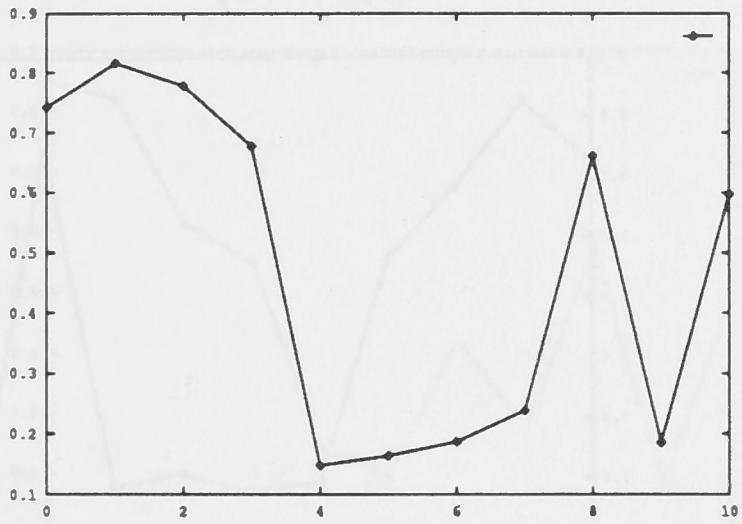
(b)



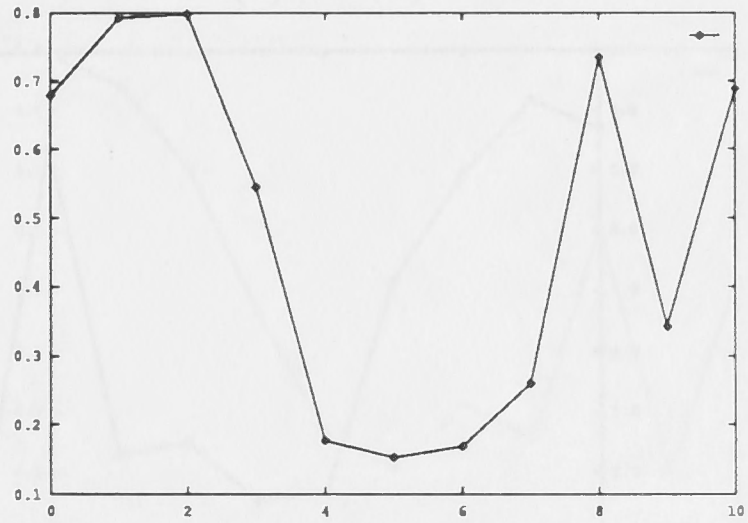
(c)



(d)



(e)



(f)

Figure A.2: *Test results of PM's* acute *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].

.

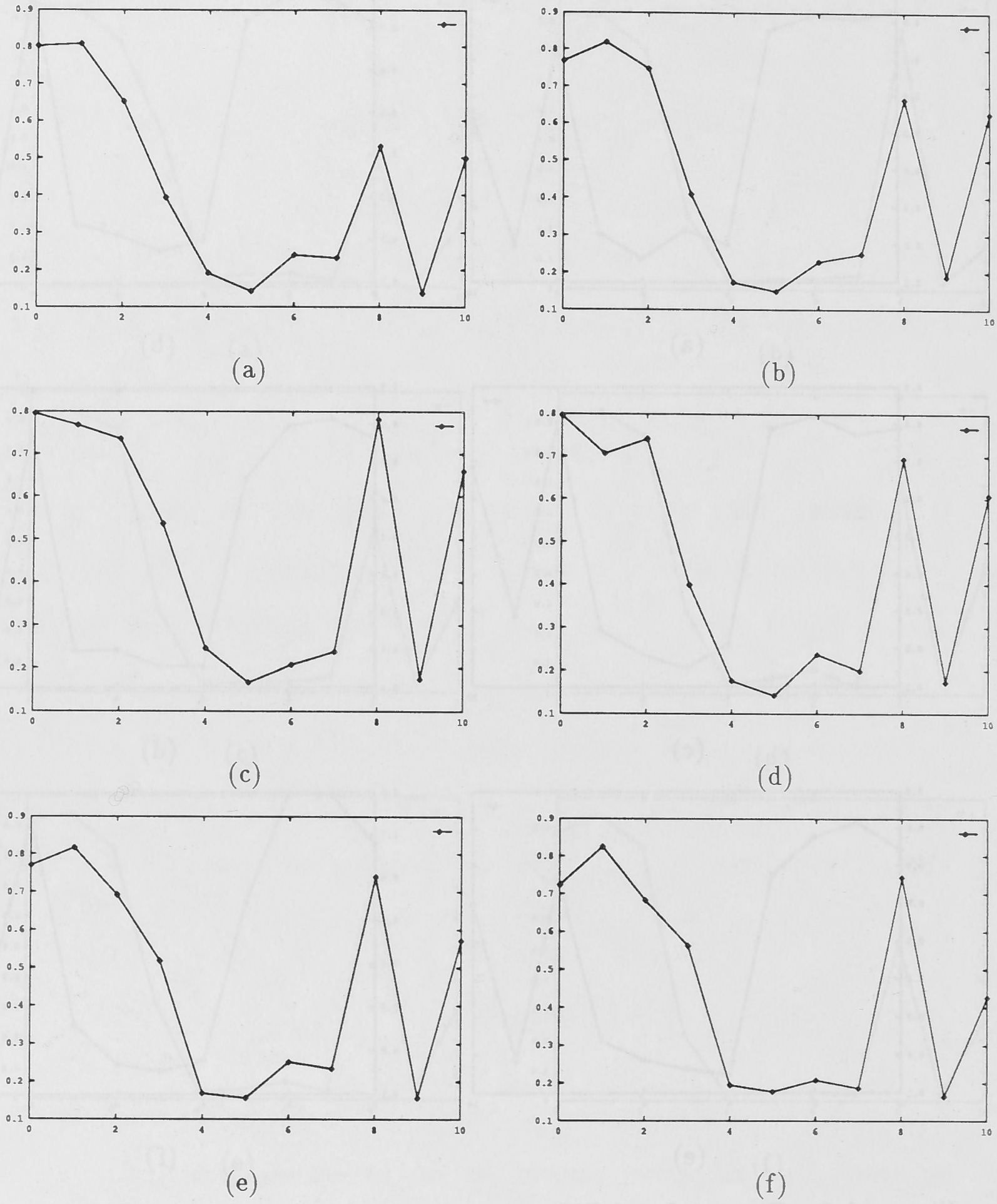
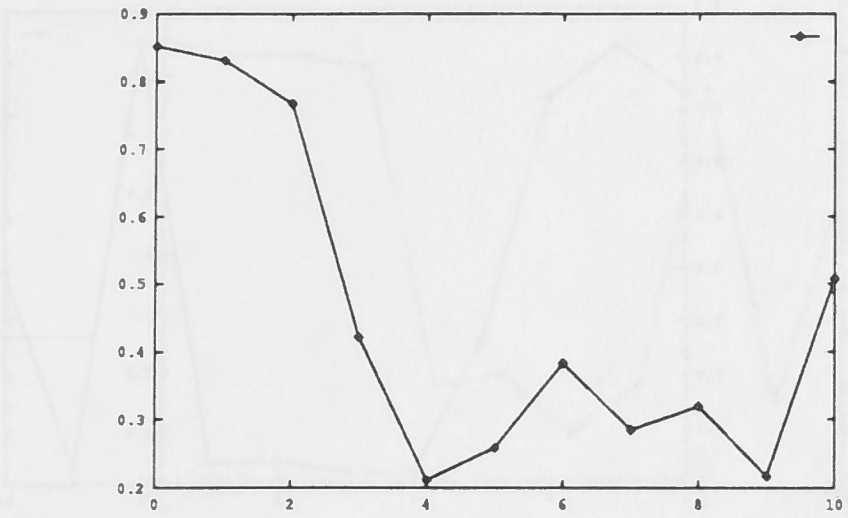
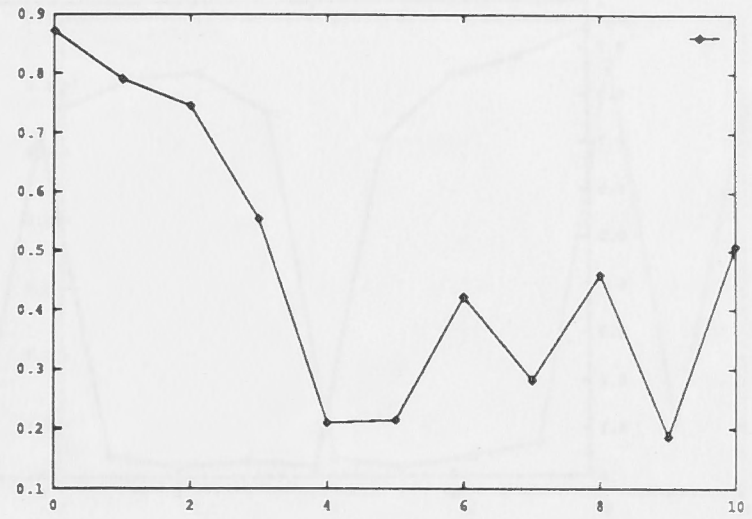
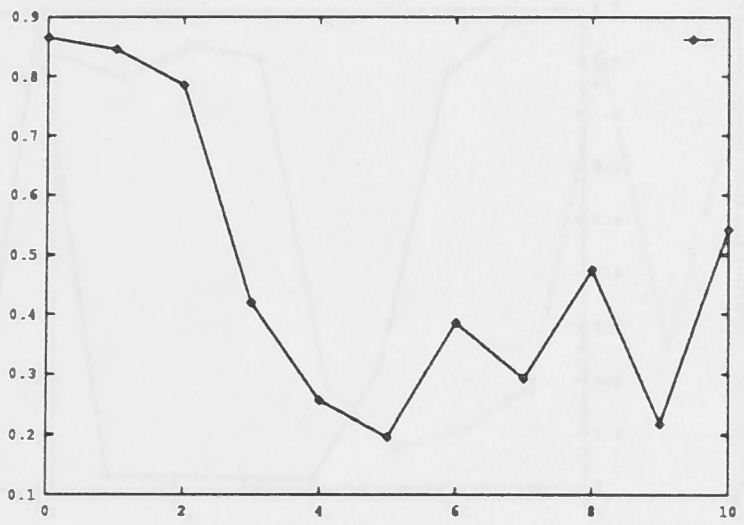
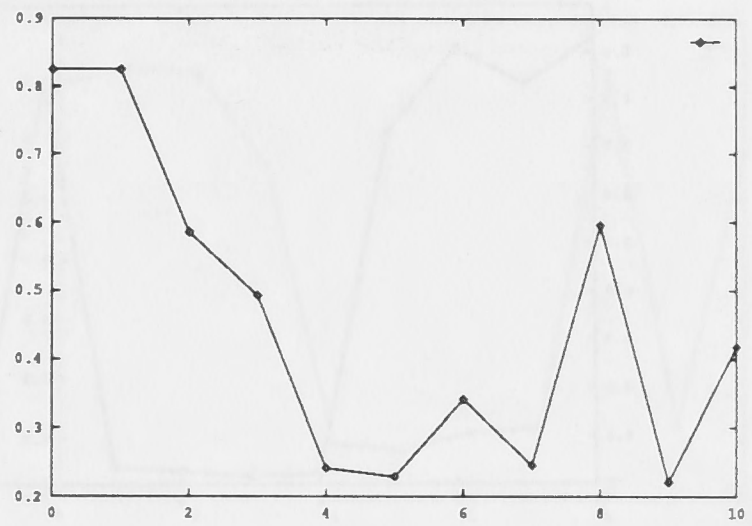


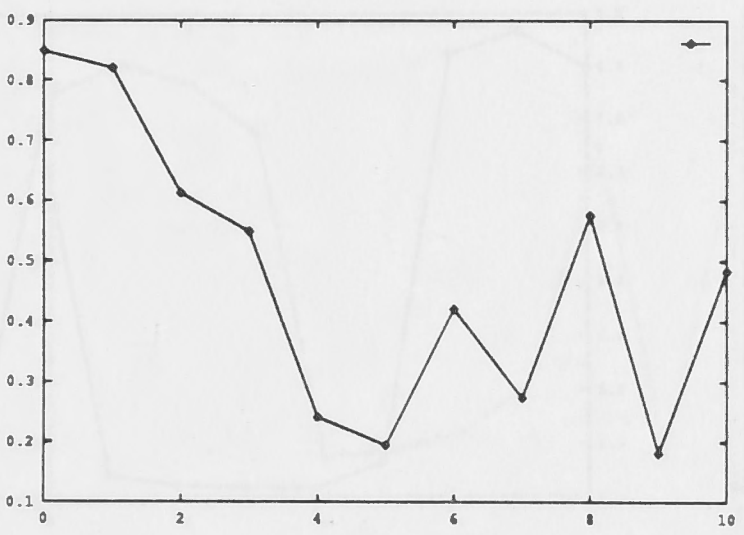
Figure A.3: *Test results of PP's* acute *feature detector tested on PP's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].











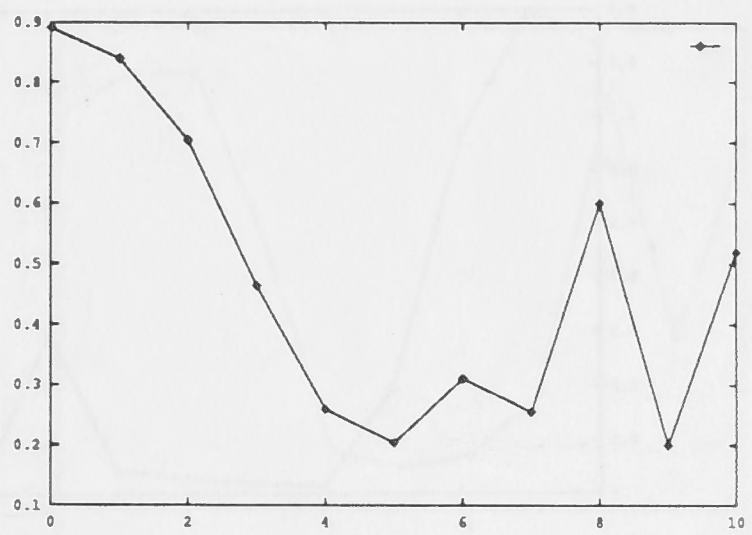
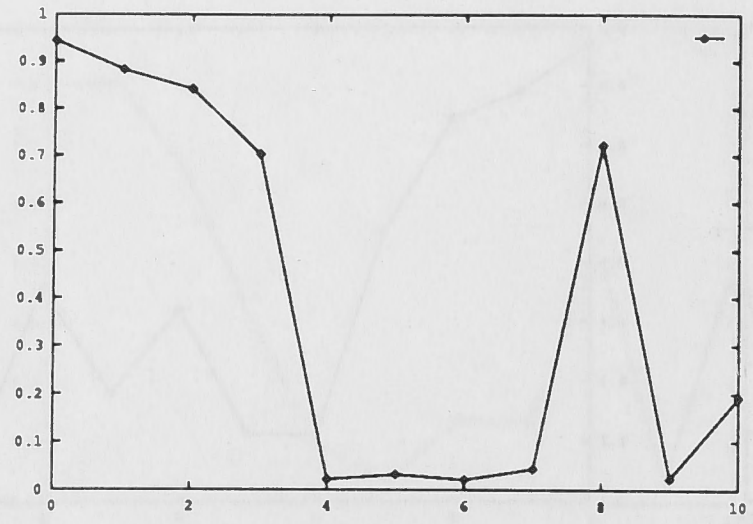
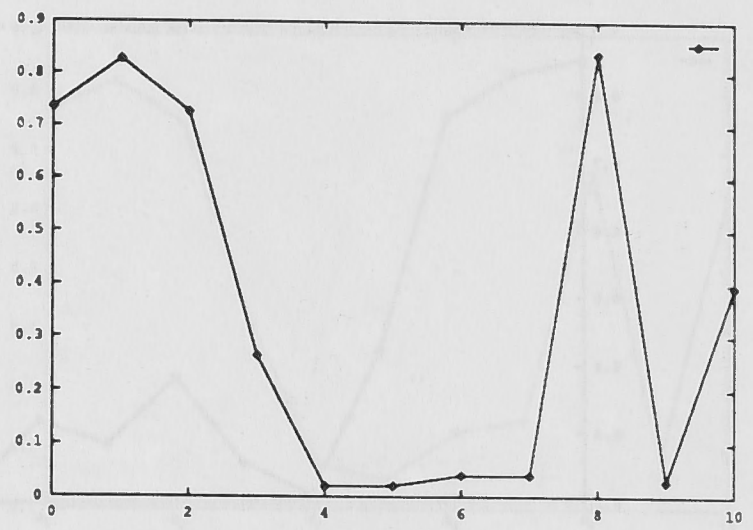
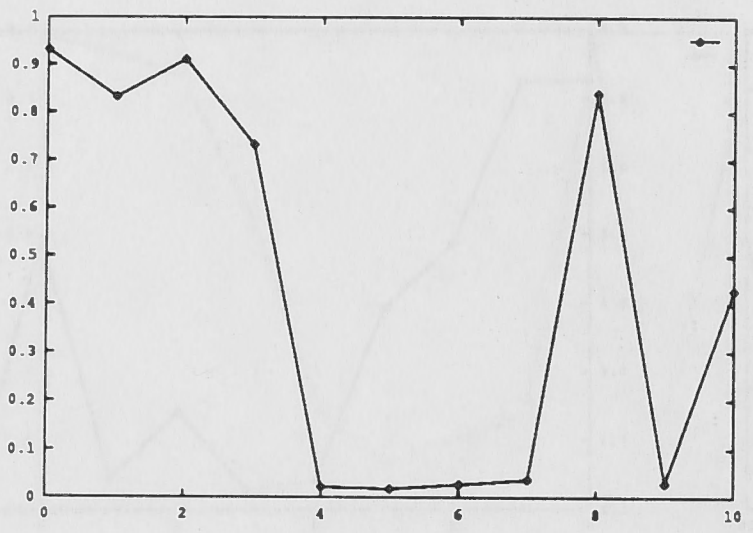
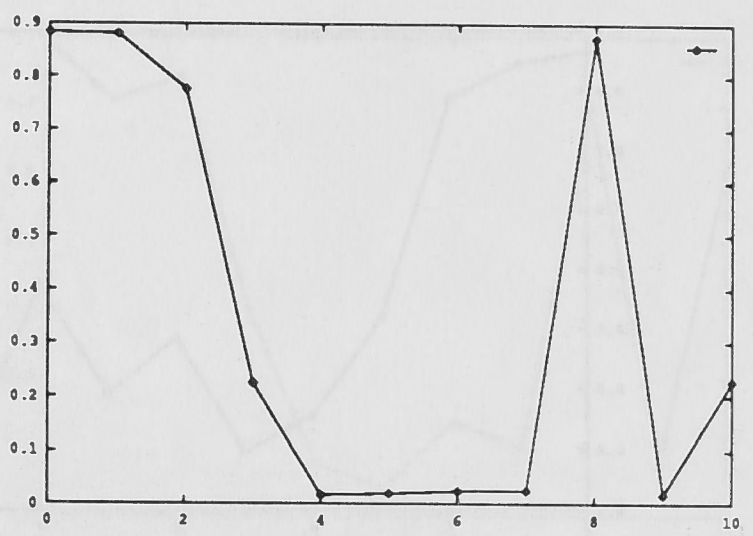


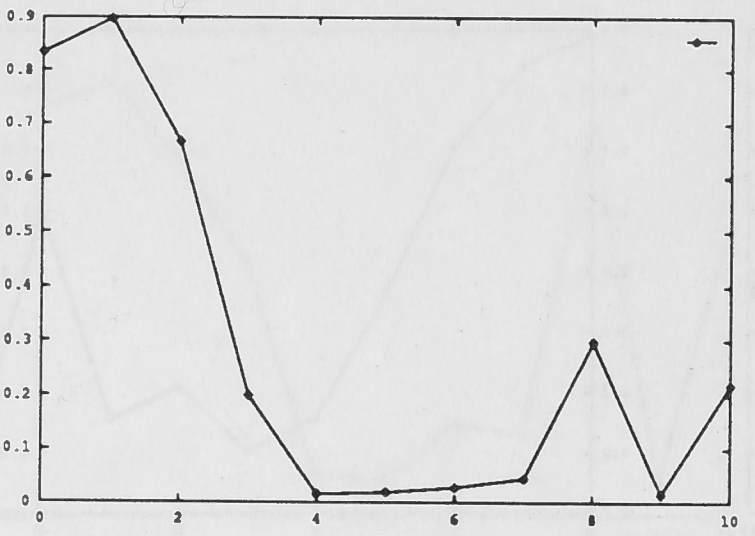
Figure A.4: *Test results of RP's* acute *feature detector tested on RP's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].











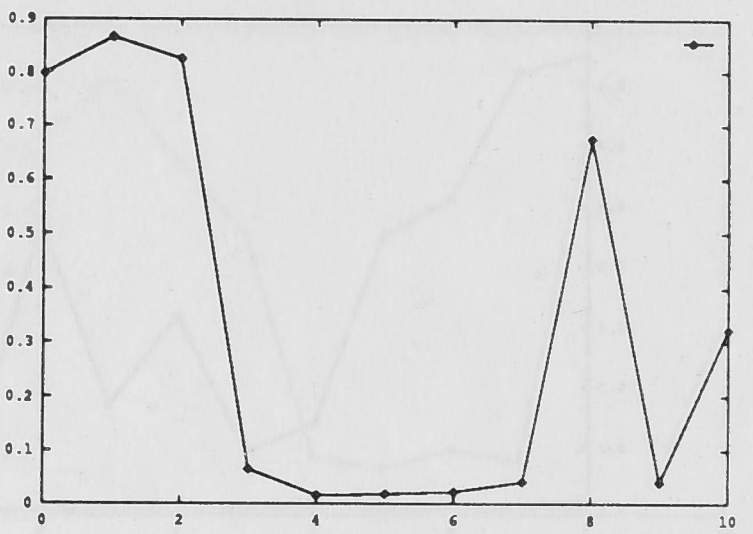
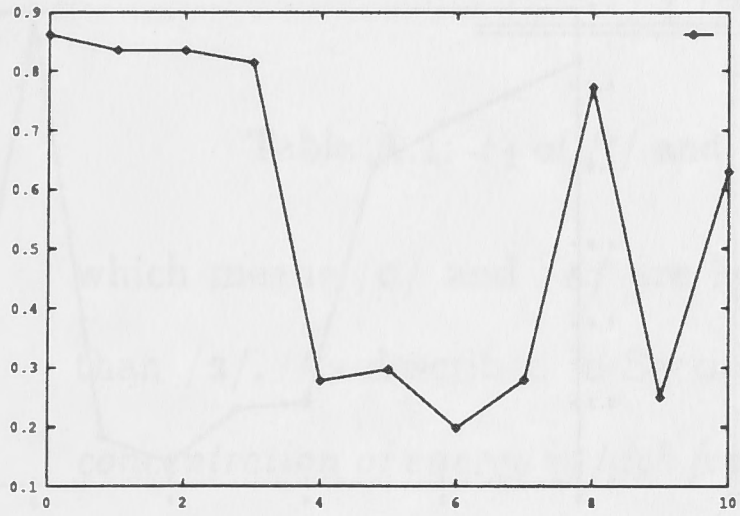
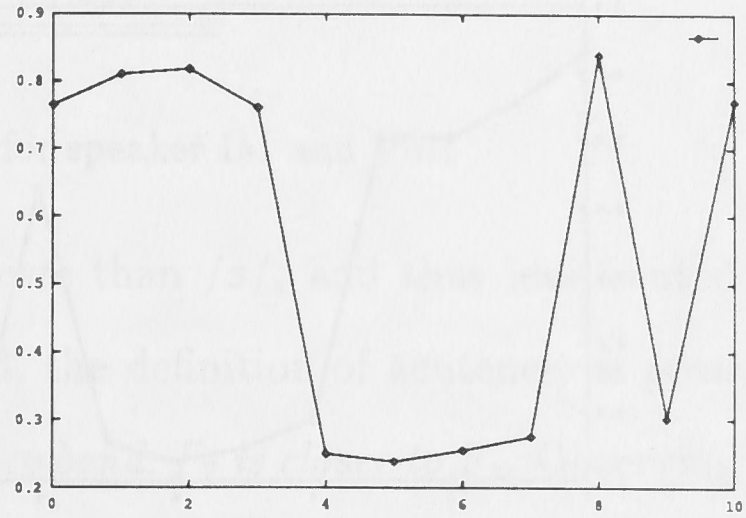


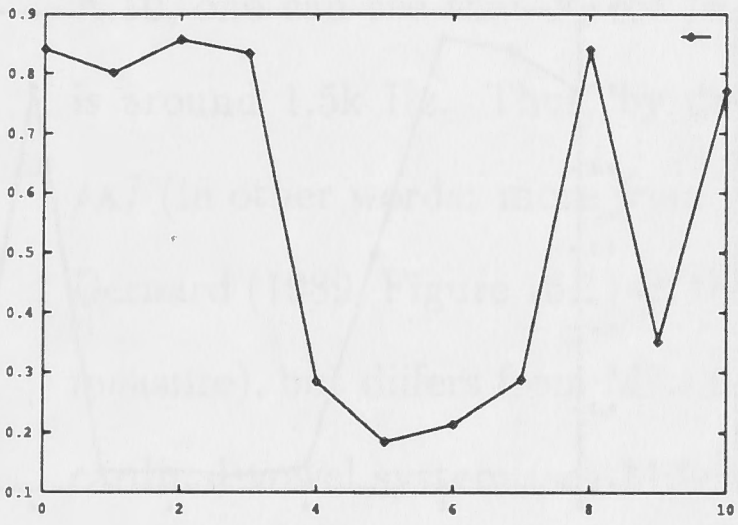
Figure A.5: *Test results of IM's* acute *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



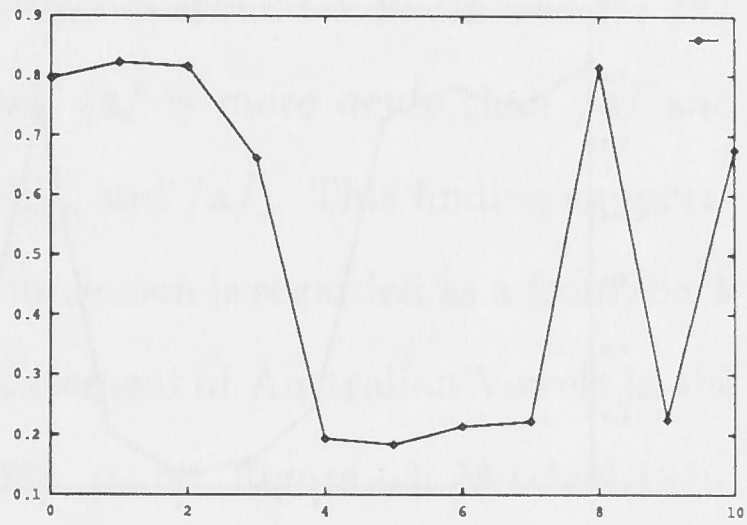
(a)



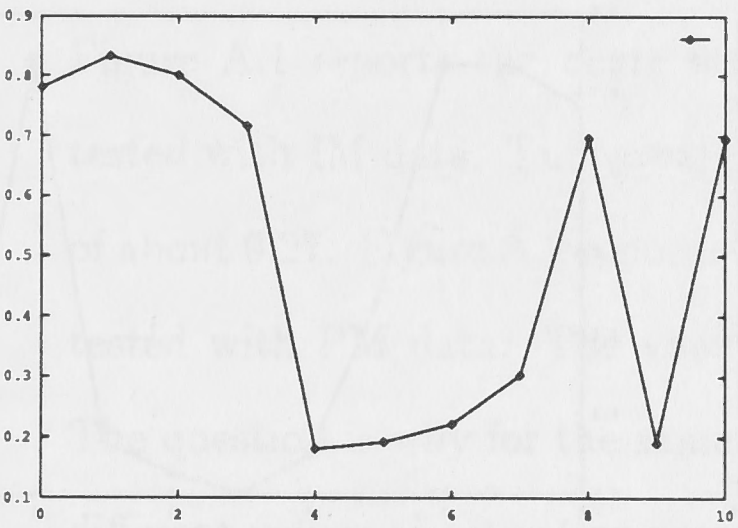
(b)



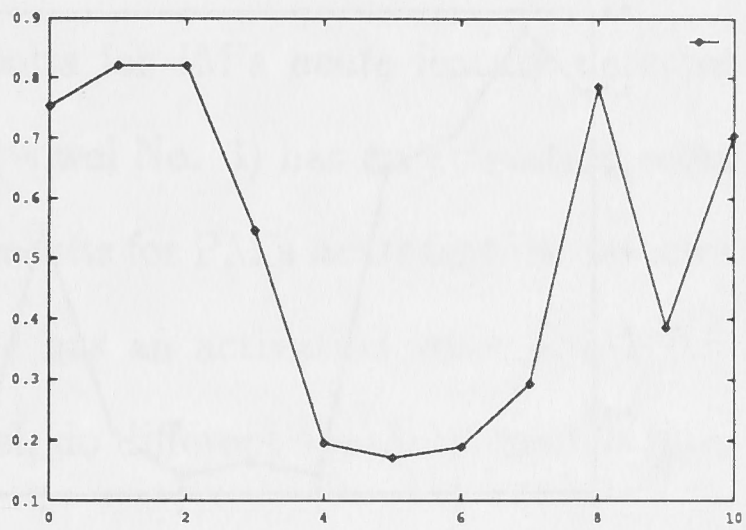
(c)



(d)

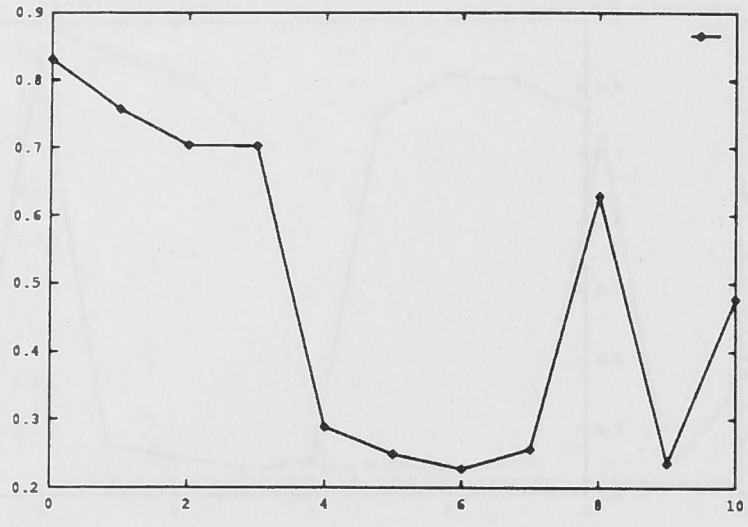


(e)

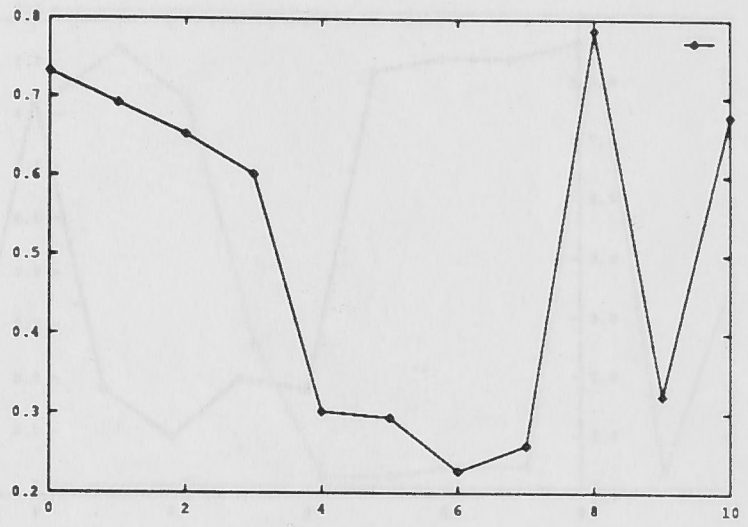


(f)

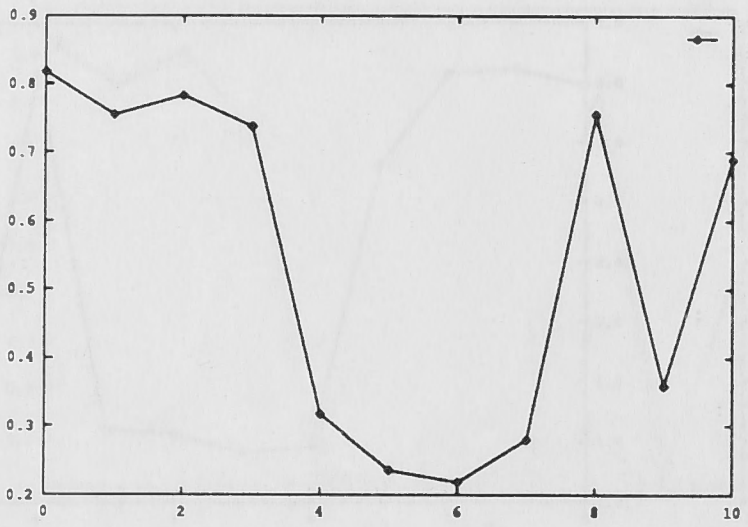
Figure A.6: *Test results of PP's* acute *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



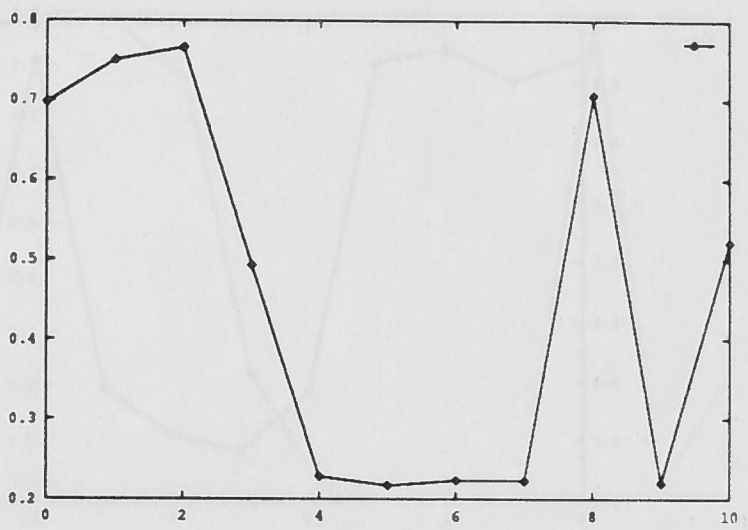
(a)



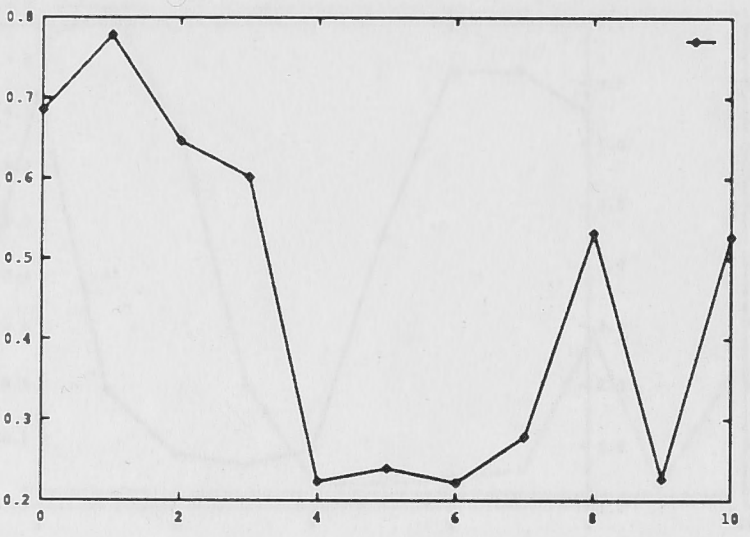
(b)



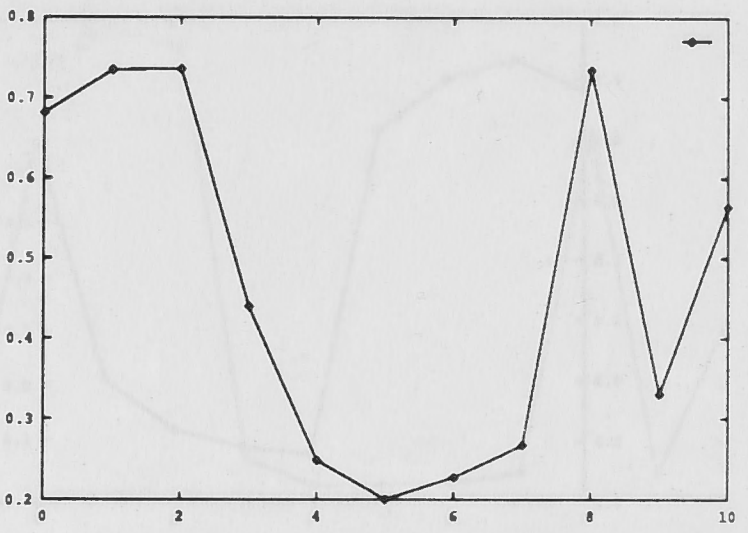
(c)



(d)



(e)



(f)

Figure A.7: *Test results of RP's* acute *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

Speaker	*Vowels*	F_2 *(Hz)*
IM	/ɪ/	2148
	/æ/	1835
PM	/ɪ/	1914
	/æ/	1835

Table A.1: F_2 of /ɪ/ and /æ/ for speaker IM and PM.

which means /ɑ/ and /ʌ/ are less *acute* than /ɜ/, and thus less fronted than /ɜ/. As described in Section 4.3, the definition of acuteness is *great concentration of energy in high frequency band, F_2 is closer to F_3*. Observing the spectrograms of [**bɑd**], [**bʌd**] and [**bɜd**] shown in Figures A.8, A.9 and A.10, one can see that F_2 for /ɑ/ and /ʌ/ is around 1.3k Hz and for /ɜ/ is around 1.5k Hz. Thus, by definition, /ɜ/ is more *acute* than /ɑ/ and /ʌ/ (in other words: more *front* than /ɑ/, and /ʌ/). This finding supports Bernard (1989, Figure 16.1) (if the F_2 dimension is regarded as a front/back measure), but differs from Mitchell's placement of Australian Vowels in the cardinal vowel system (see Mitchell 1962, p. 63, Figure 63; Mitchell 1946, p. 30, Figure 1; Mitchell and Delbridge 1965, p. 35, Figure 1).

- Figure A.1 reports the *acute* test results for IM's *acute* feature detector tested with IM data. The vowel /æ/ (vowel No. 3) has an activation score of about 0.27. Figure A.2 reports test results for PM's *acute* feature detector tested with PM data. The vowel /æ/ has an activation score above 0.5. The question is why for the same vowel, do different speakers' models have different values of activation scores?

  Analysing the spectrograms for /ɪ/ and /æ/ in [bVd] context for IM and PM using XWAVES, (see Figures A.11 through A.14), the F_2 data obtained for these two sounds for the given speakers are summarised in Table A.1.

  /ɪ/ is one of the three reference vowel sounds for each of the speaker models. The reason for /æ/ having lower *acute* activation scores for IM than for PM

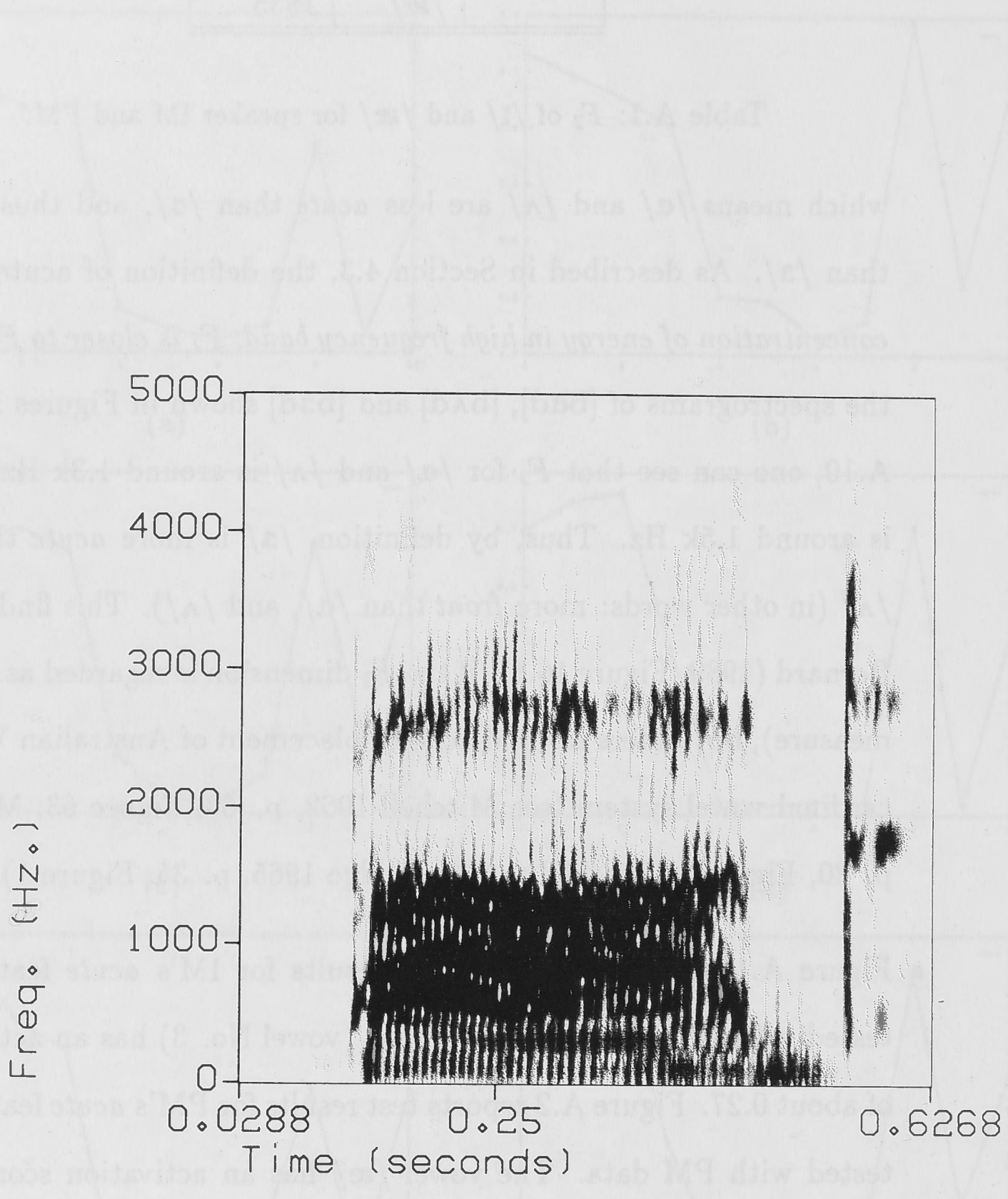


Figure A.8: Spectrogram for [**bad**] from speaker IM.

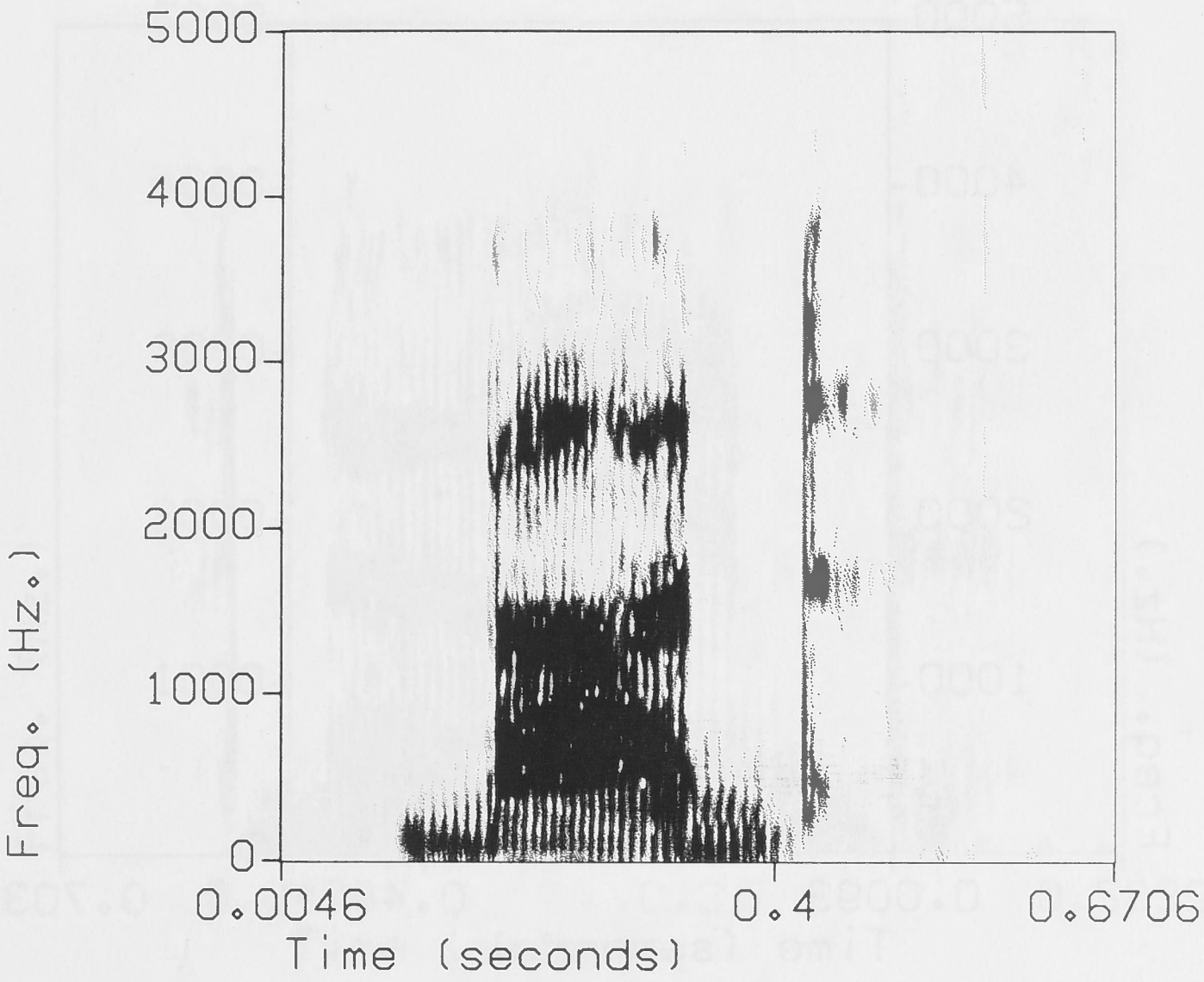


Figure A.9: Spectrogram for [**bʌd**] from speaker IM.

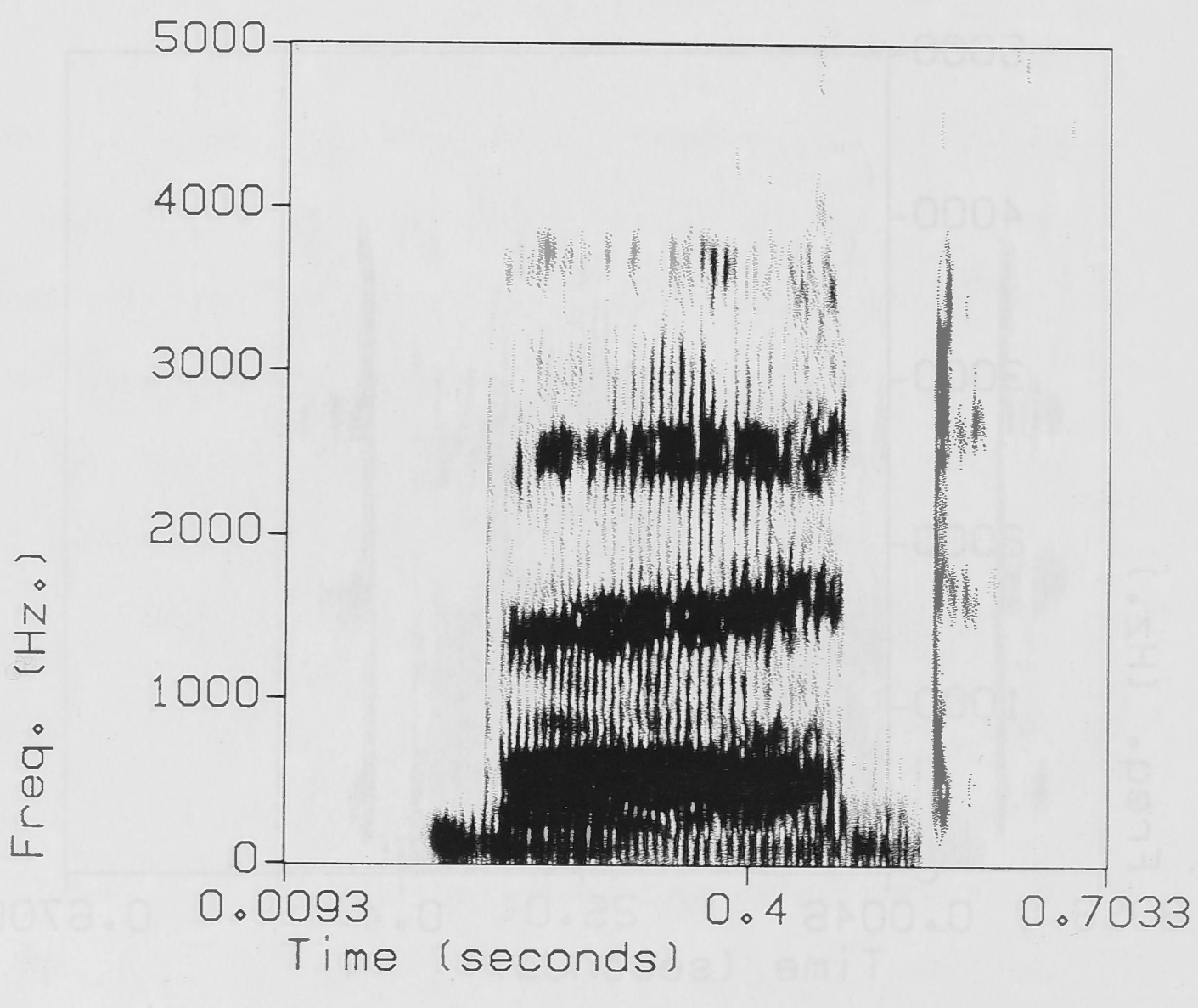


Figure A.10: Spectrogram [**b3d**] from speaker IM.

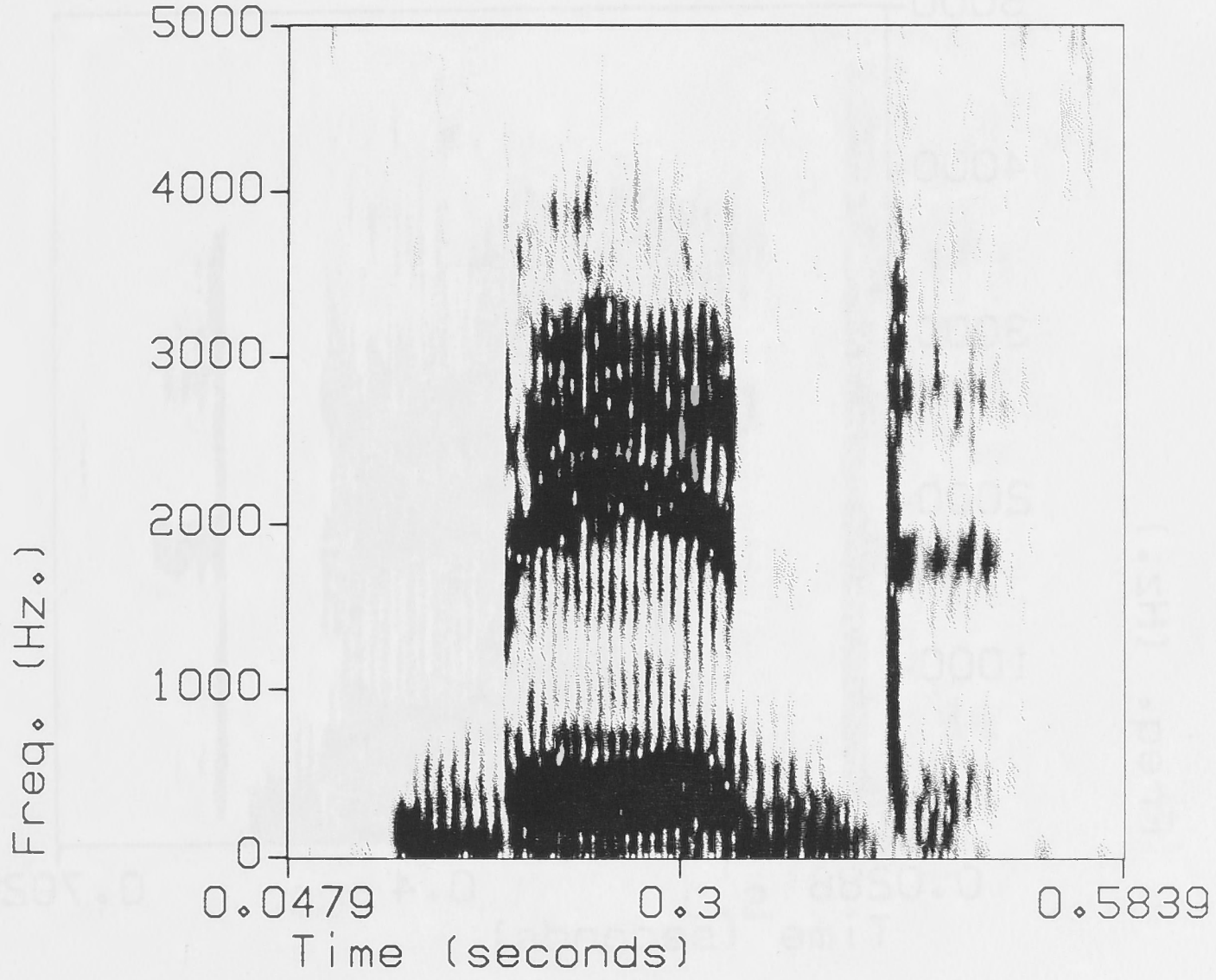


Figure A.11: Spectrogram [**bɪd**] from speaker IM.

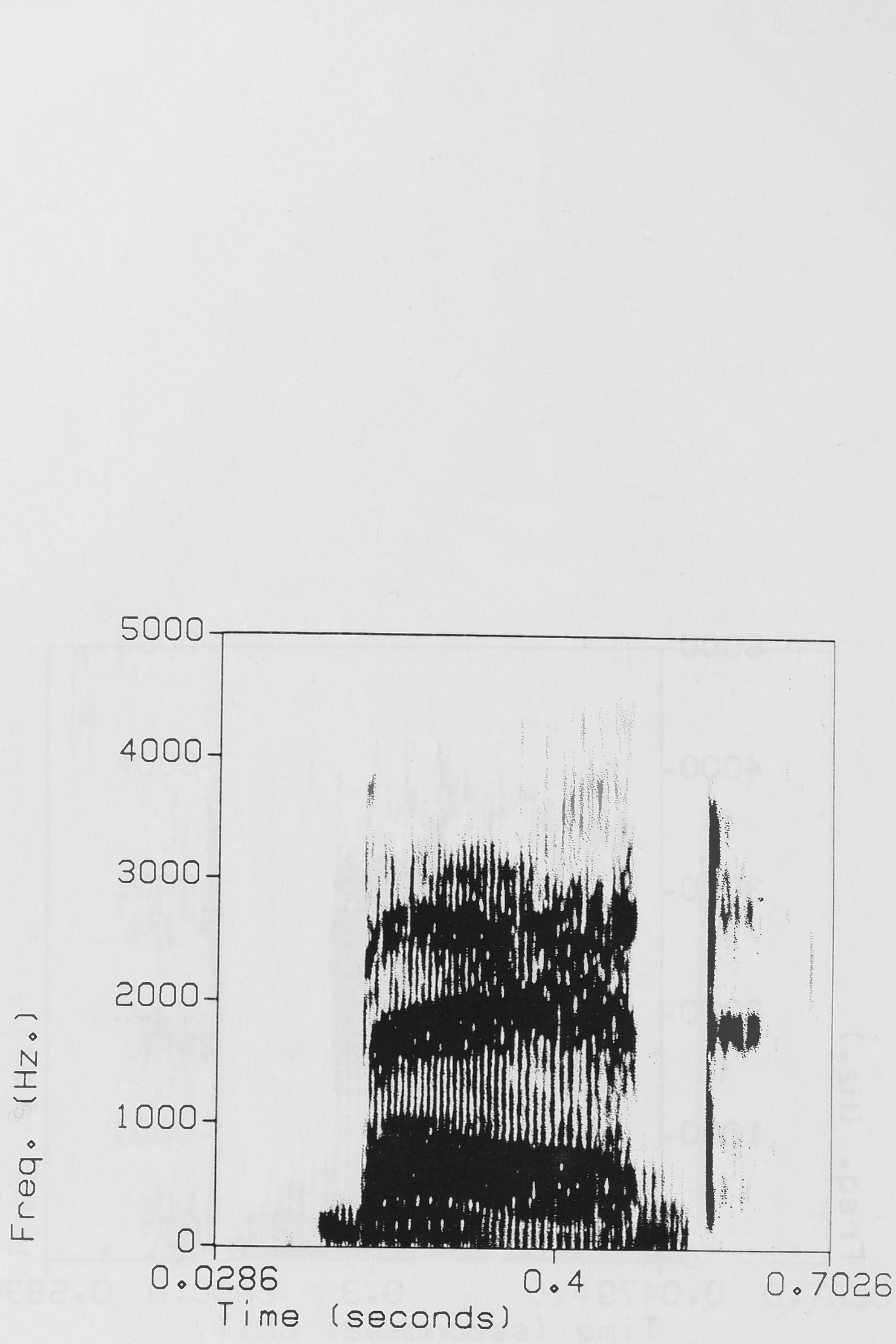


Figure A.12: Spectrogram [**bæd**] from speaker IM.

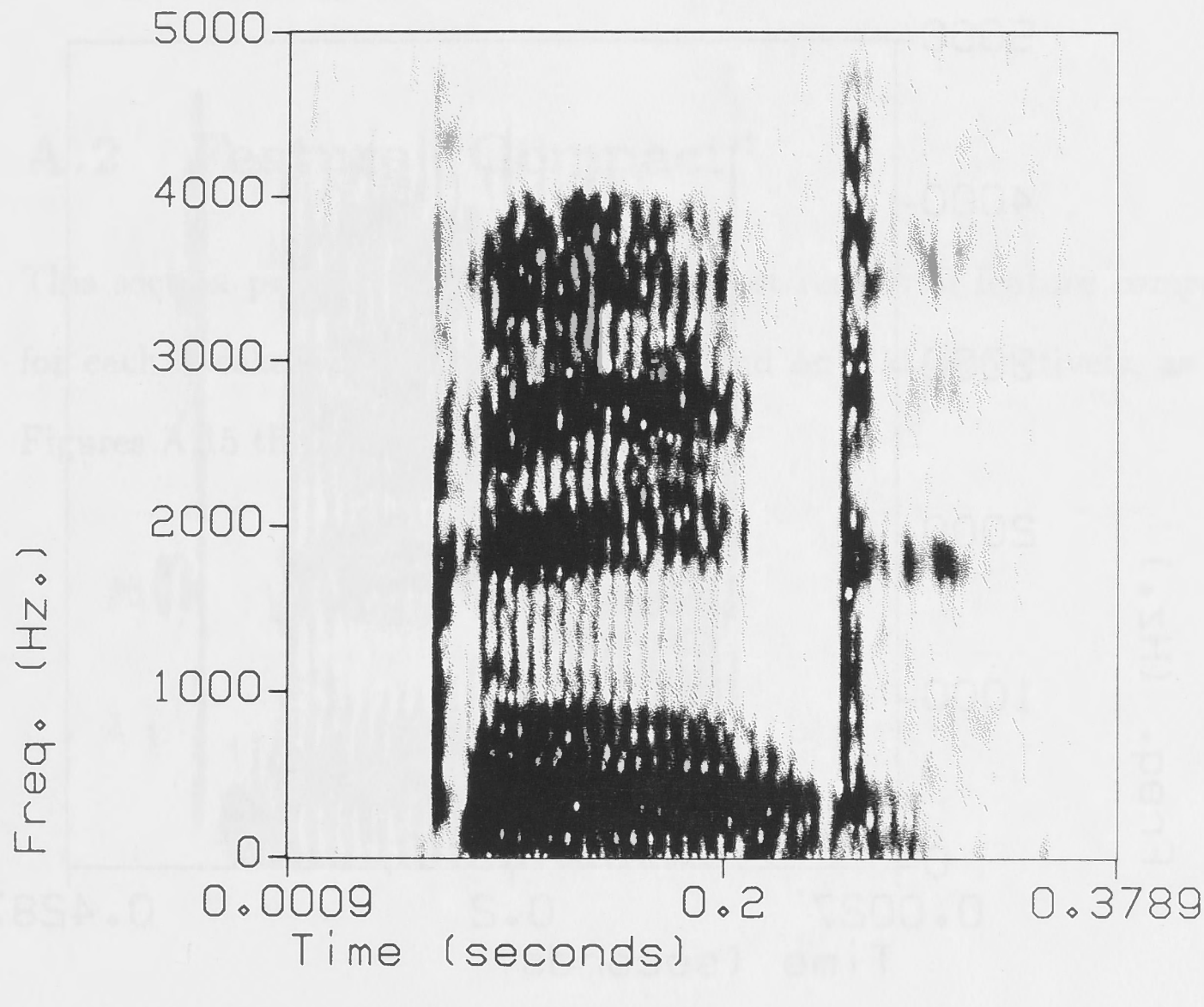


Figure A.13: Spectrogram [**bɪd**] from speaker PM.

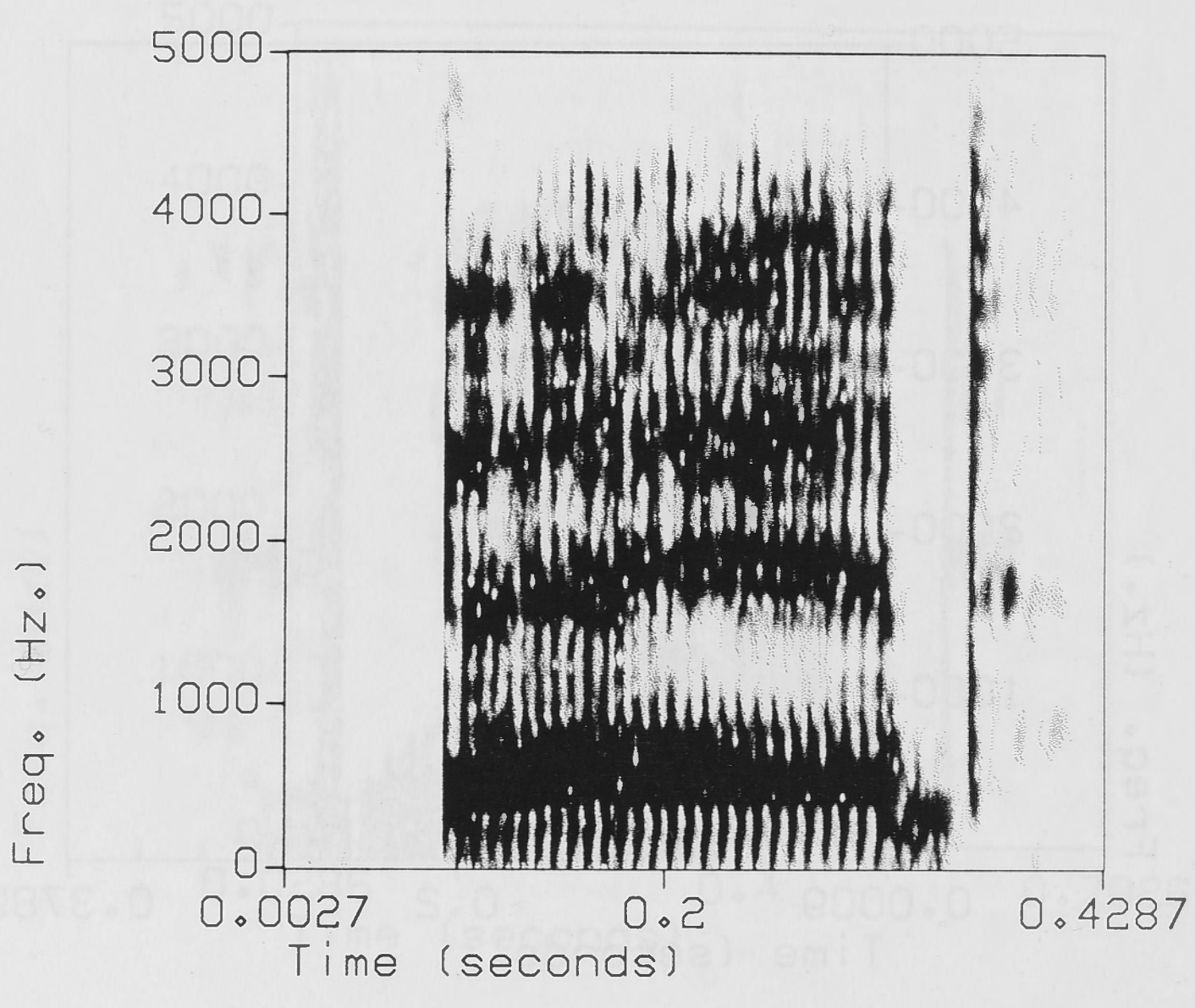


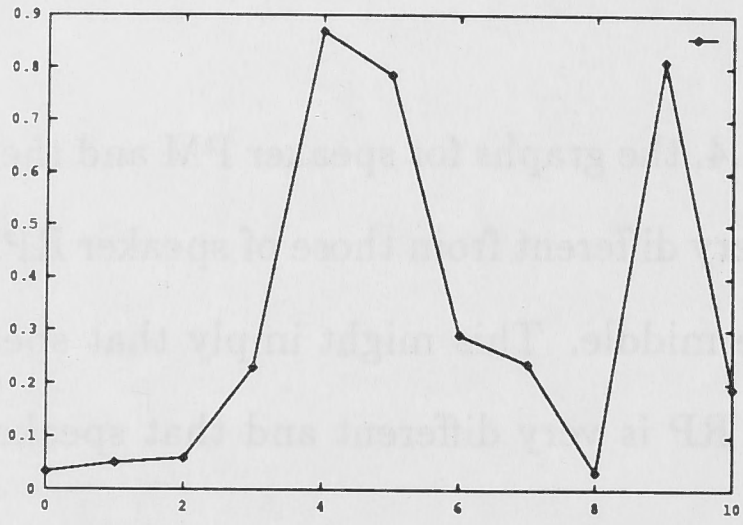
Figure A.14: Spectrogram [**bæd**] from speaker PM.

is because the reference sound for *acute* for IM is much higher than for PM, and the /æ/ itself is similar for both speakers, therefore /æ/ would be lower when there is a very high reference point (IM case) than when there is a low reference point (PM case).

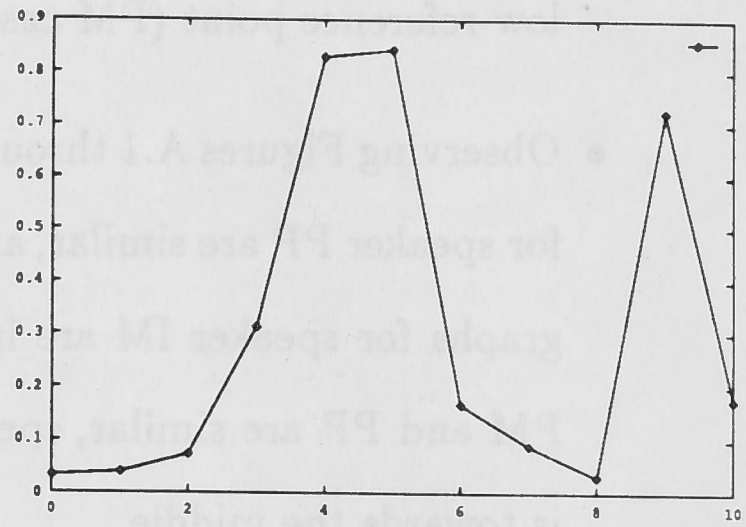
- Observing Figures A.1 through A.4, the graphs for speaker PM and the ones for speaker PP are similar, and very different from those of speaker RP. The graphs for speaker IM are in the middle. This might imply that speakers PM and PP are similar, speaker RP is very different and that speaker IM is towards the middle.

A.2 Feature "Compact"

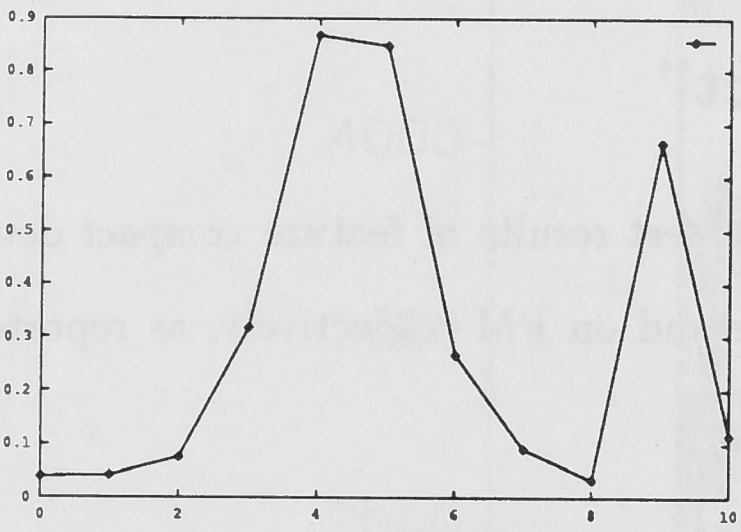
This section presents a complete set of test results of feature *compact* detector for each speaker, tested on themselves and on PM respectively, as reported in Figures A.15 through A.21.



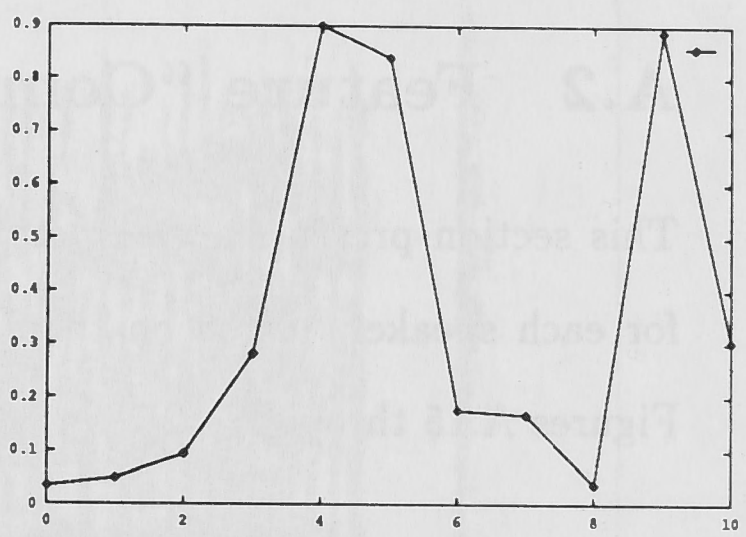
(a)



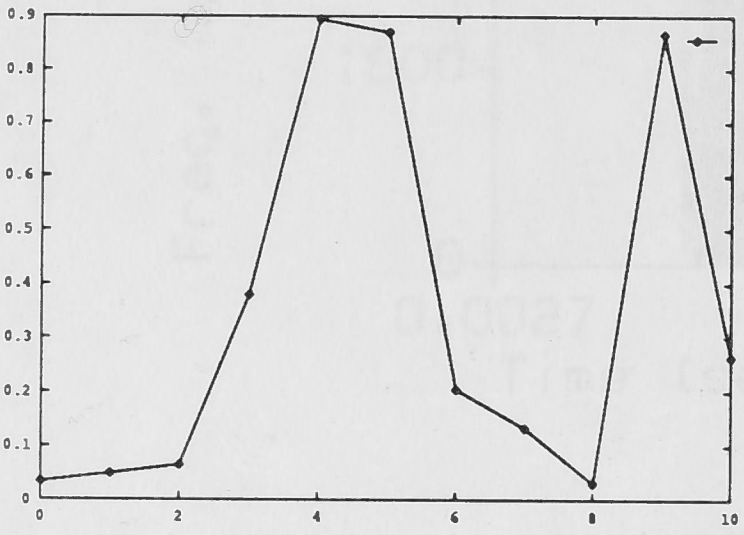
(b)



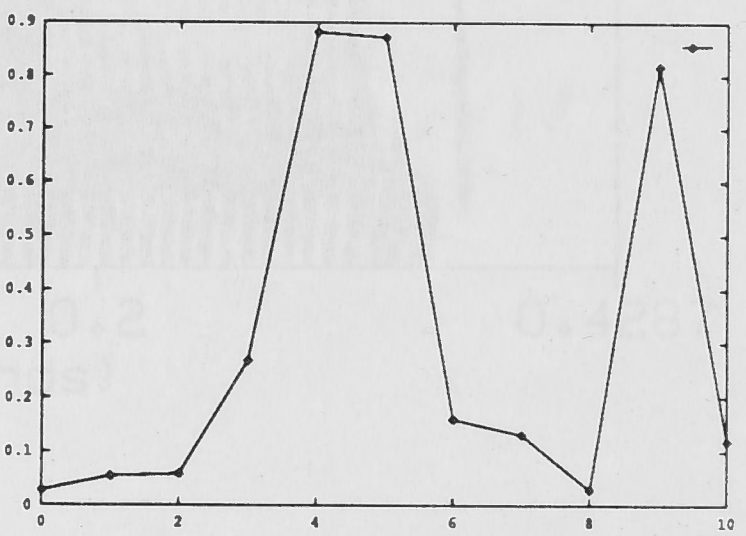
(c)



(d)

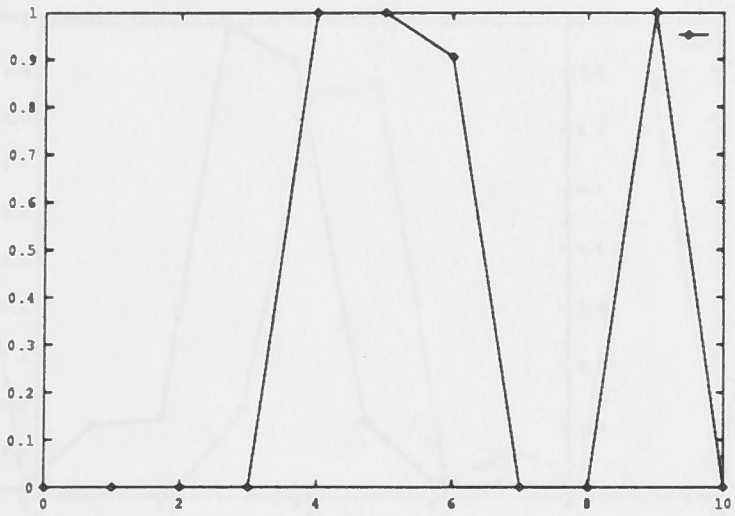


(e)

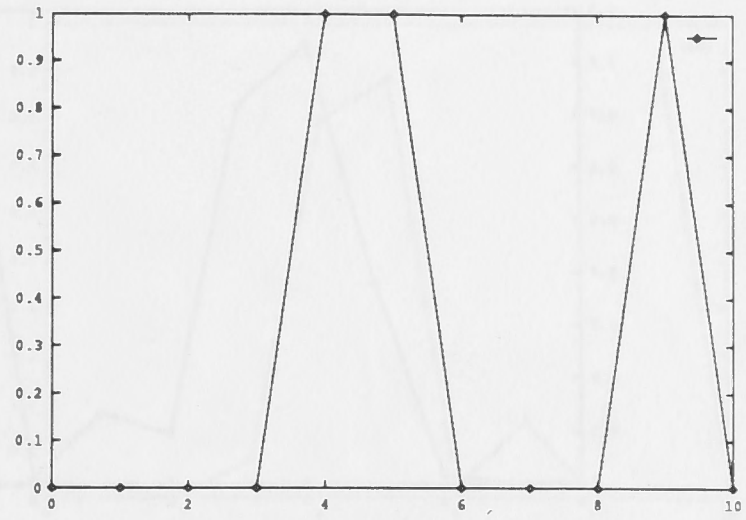


(f)

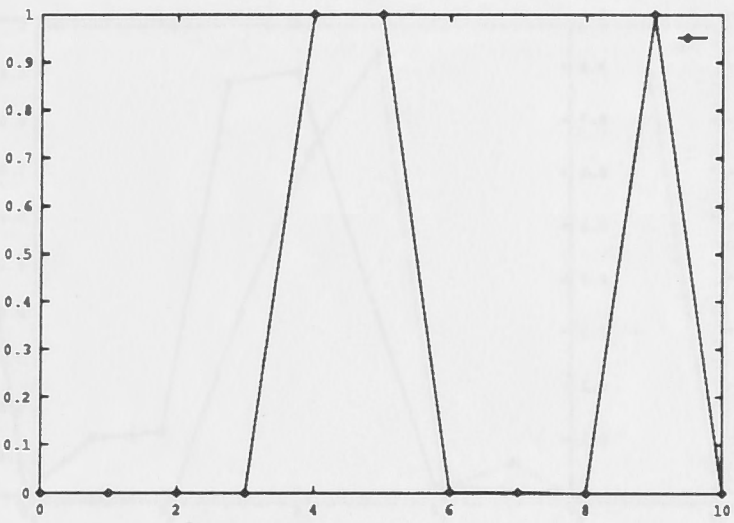
Figure A.15: *Test results of IM's* compact *feature detector tested on IM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



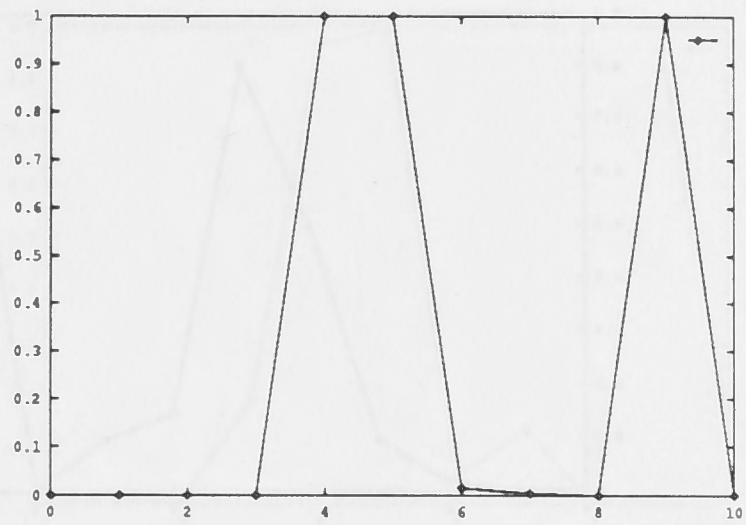
(a)



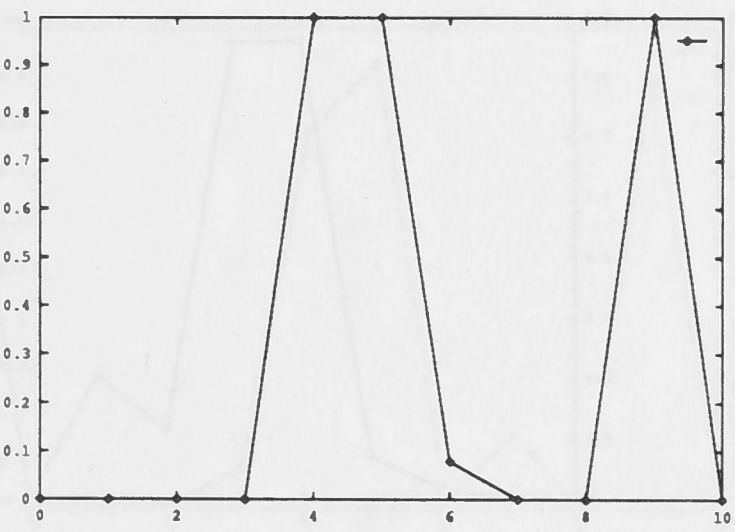
(b)



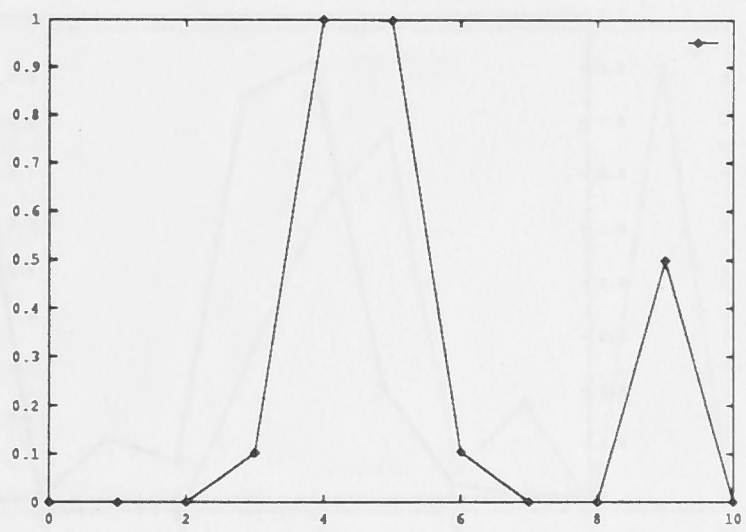
(c)



(d)

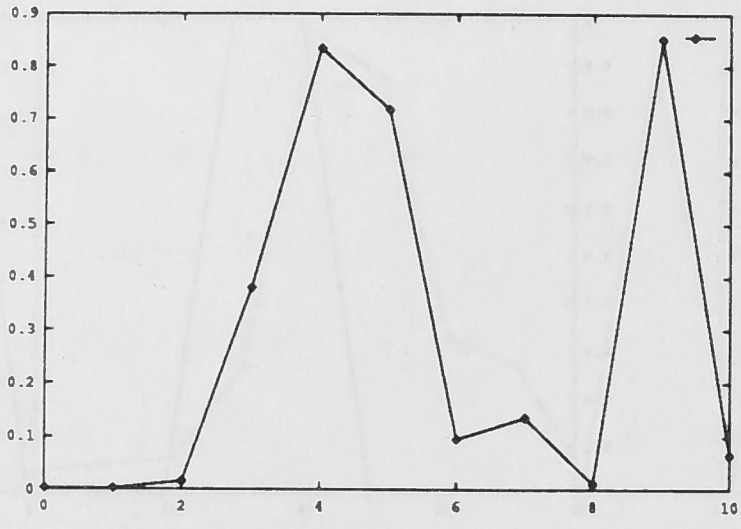


(e)

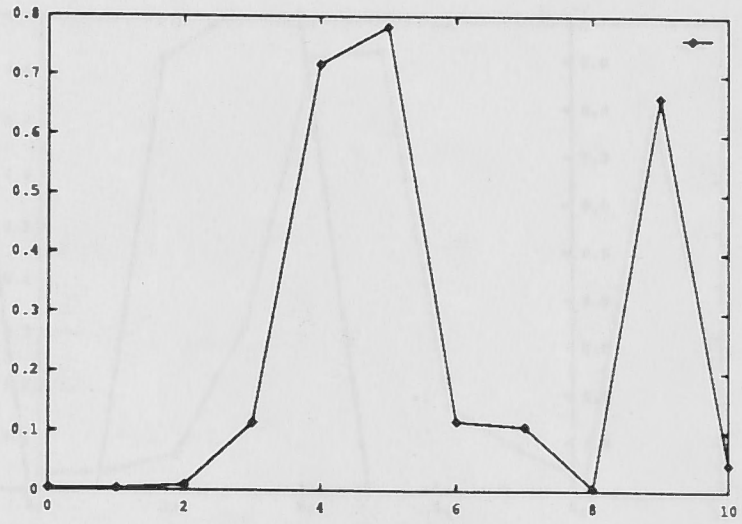


(f)

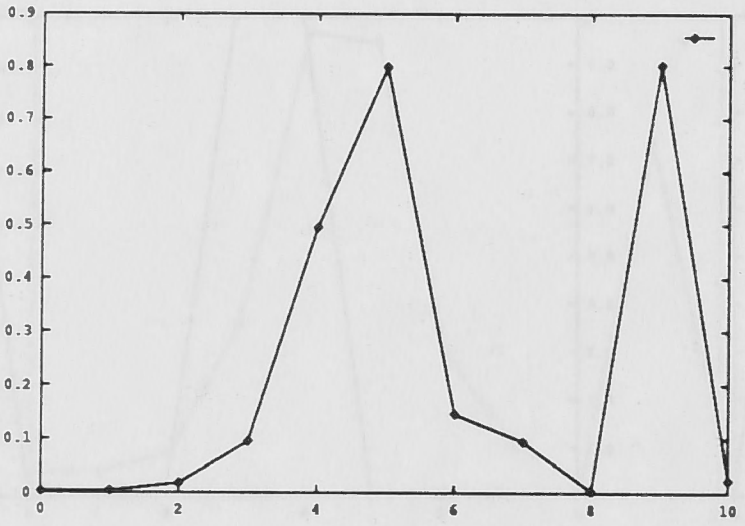
Figure A.16: *Test results of PM's* compact *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



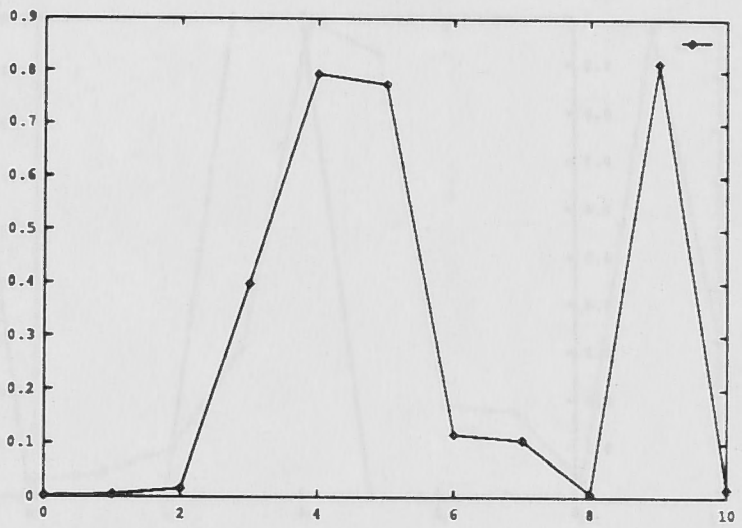
(a)



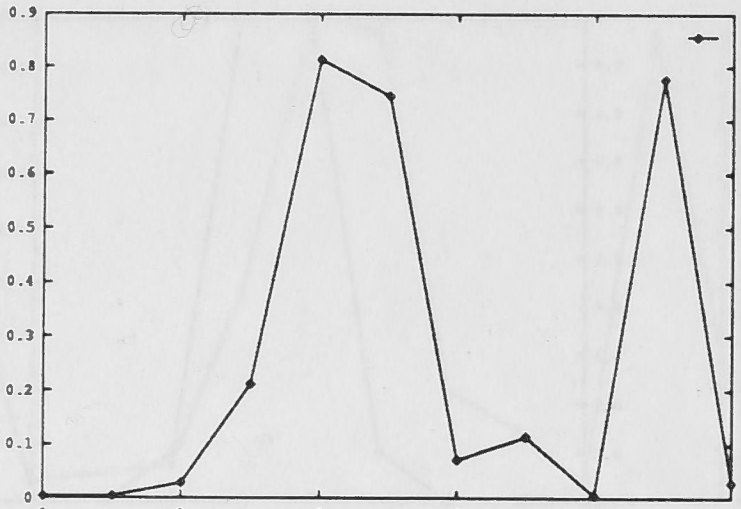
(b)



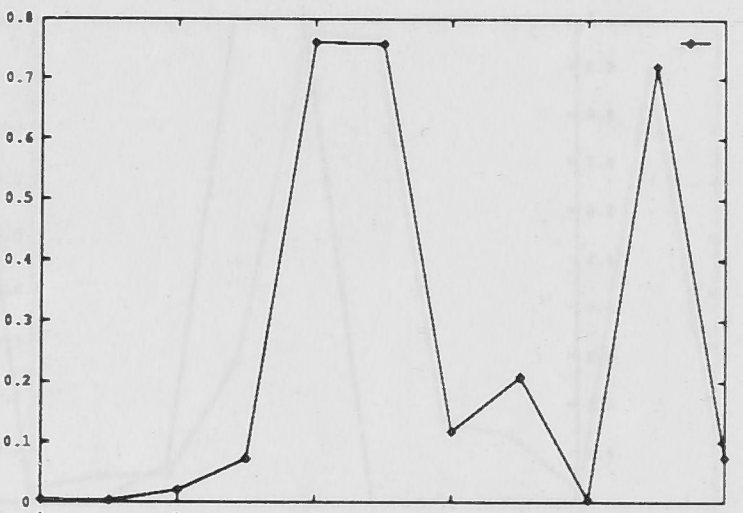
(c)



(d)

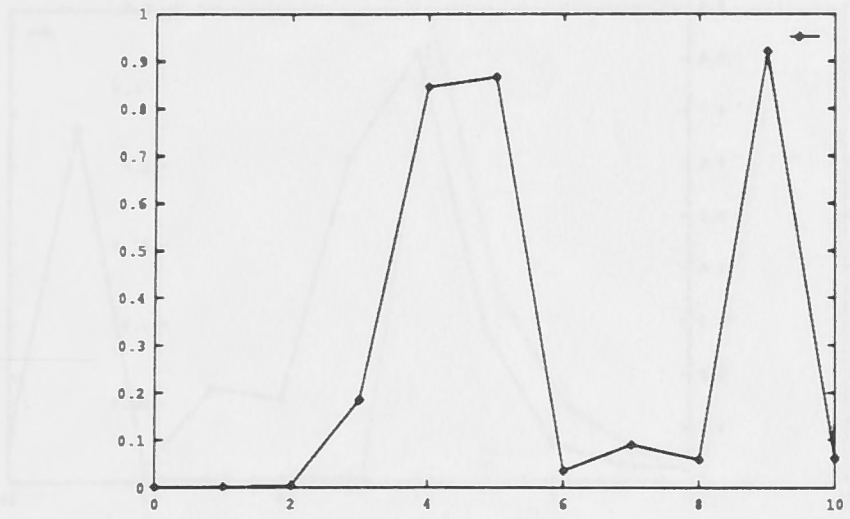


(e)

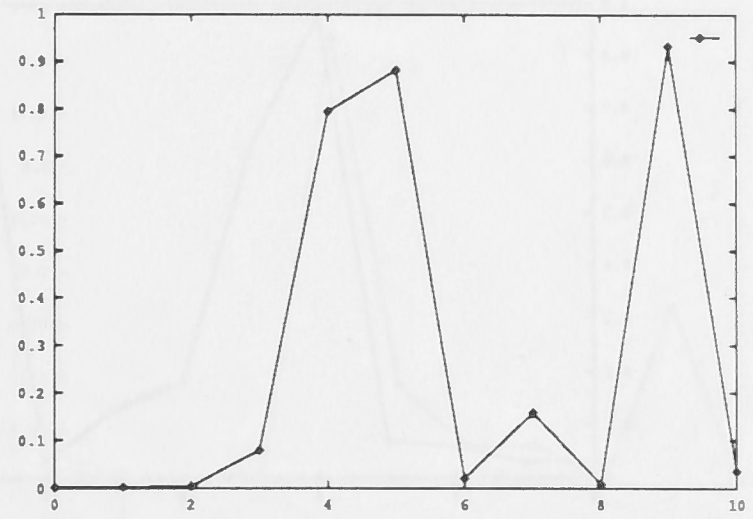


(f)

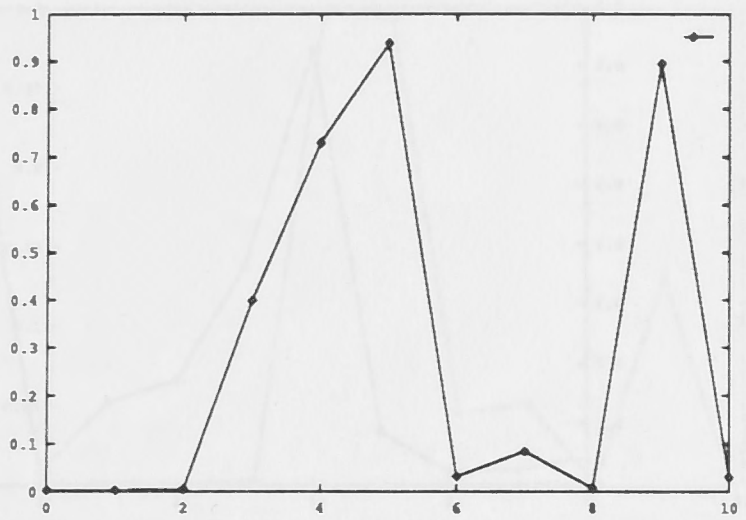
Figure A.17: *Test results of PP's* compact *feature detector tested on PP's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



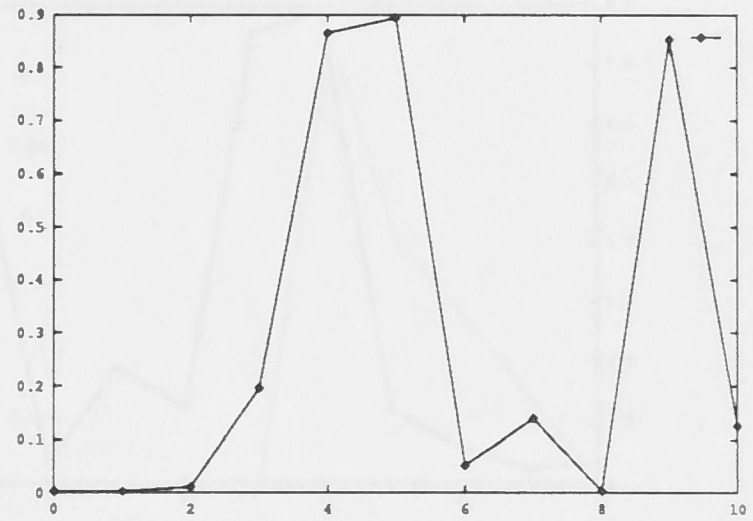
(a)



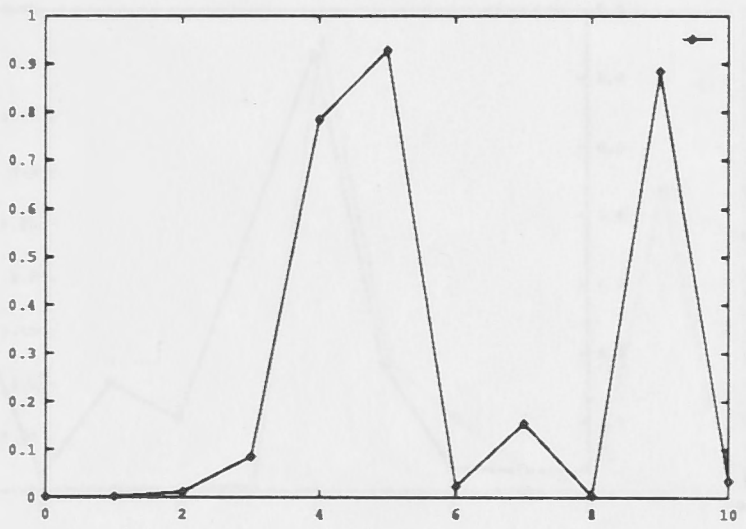
(b)



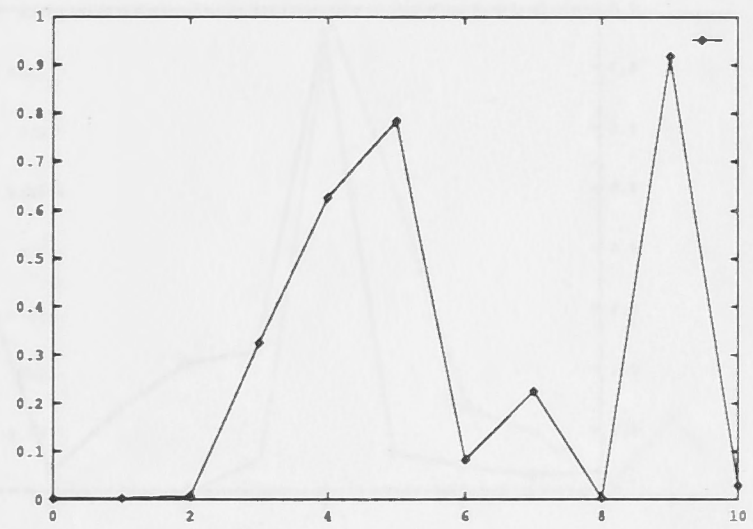
(c)



(d)

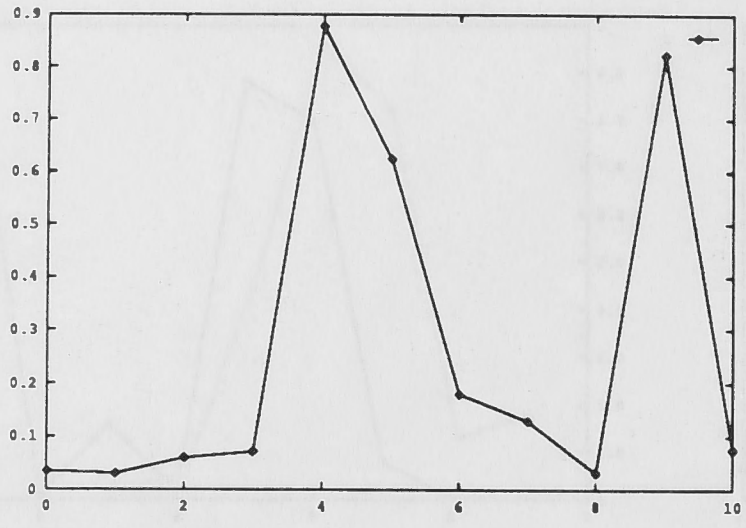


(e)

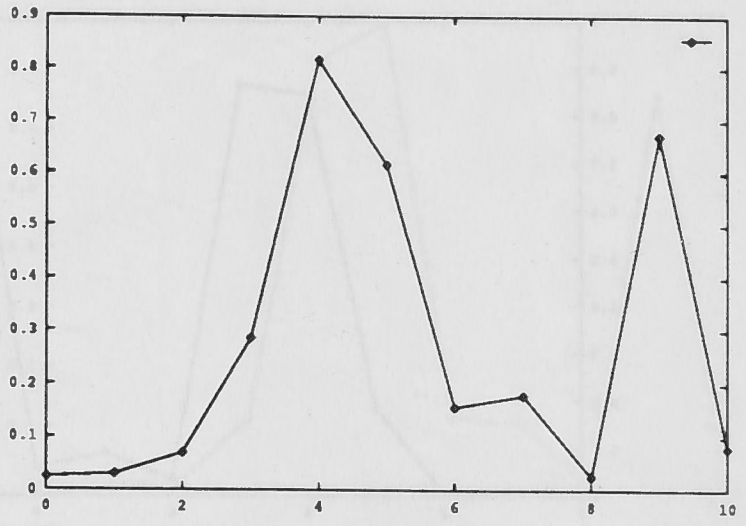


(f)

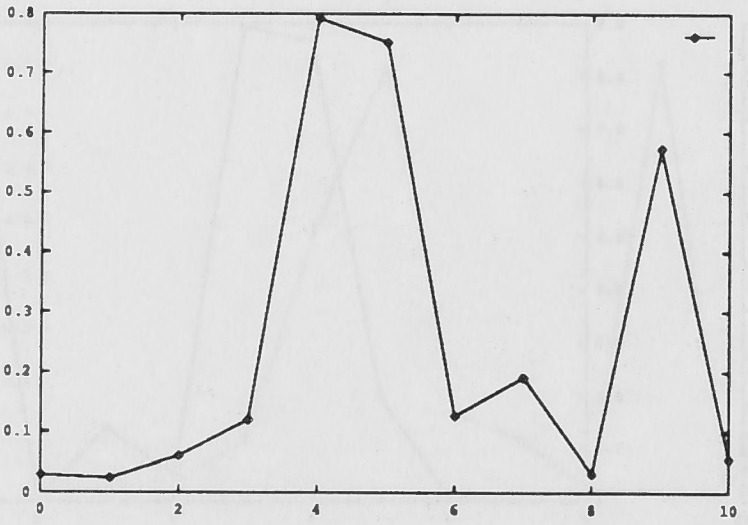
Figure A.18: *Test results of RP's* compact *feature detector tested on RP's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



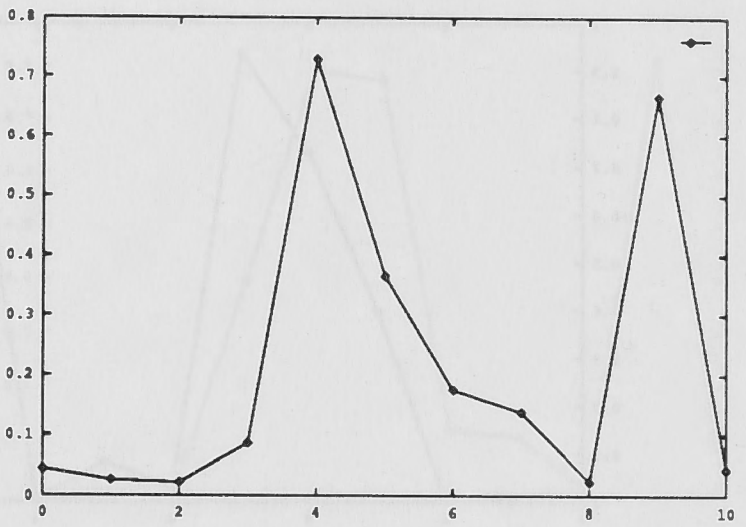
(a)



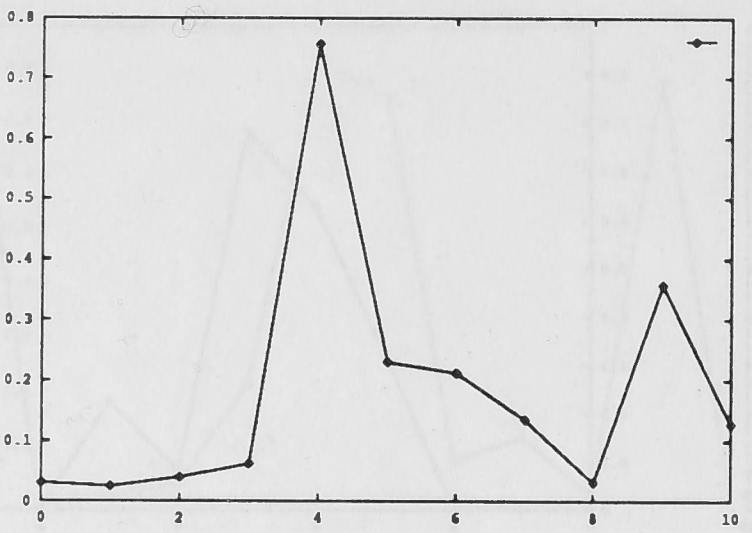
(b)



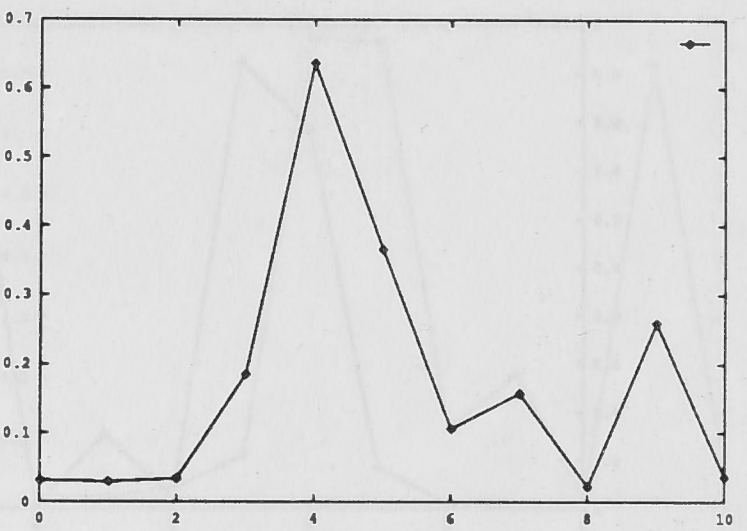
(c)



(d)

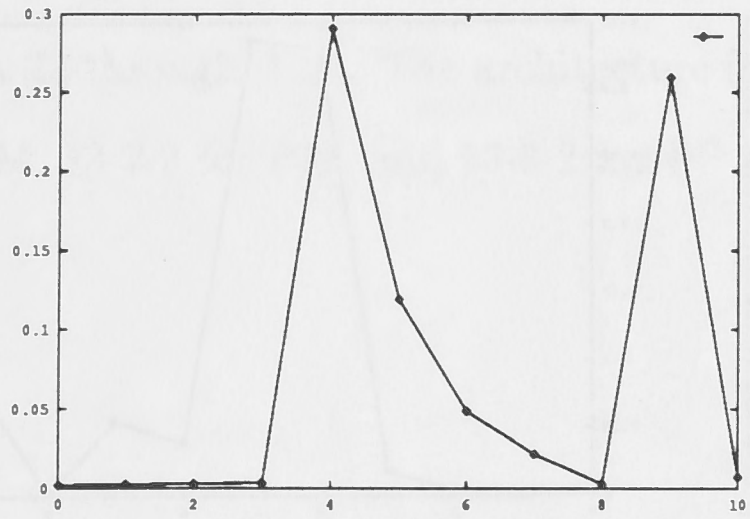


(e)

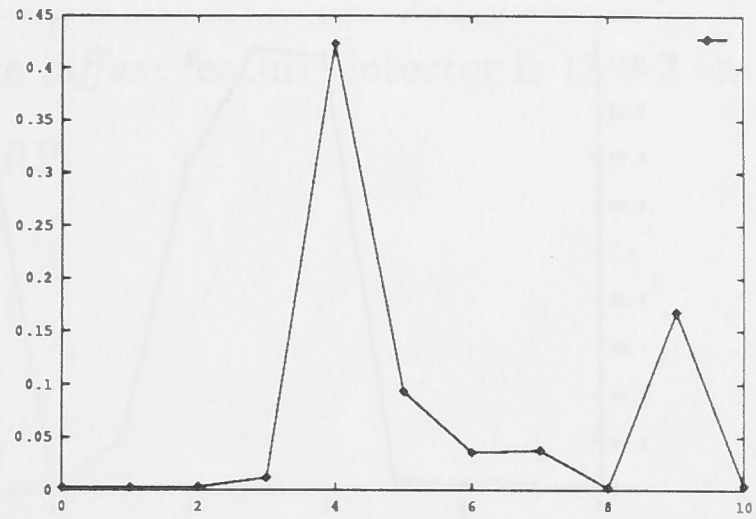


(f)

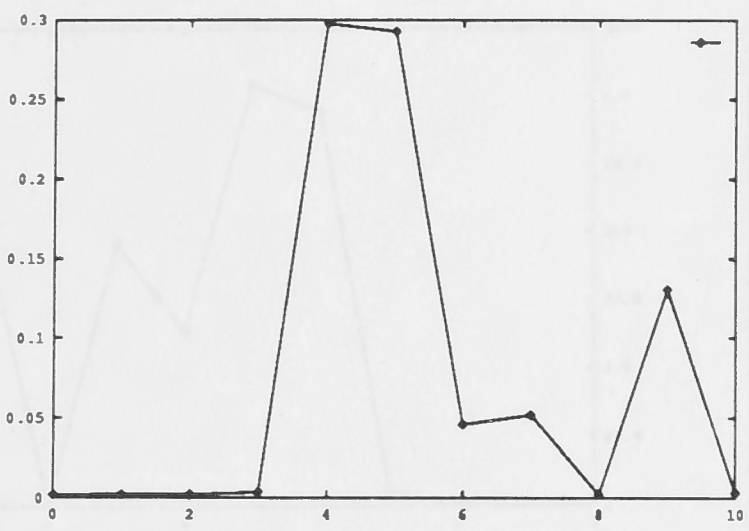
Figure A.19: *Test results of IM's* compact *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



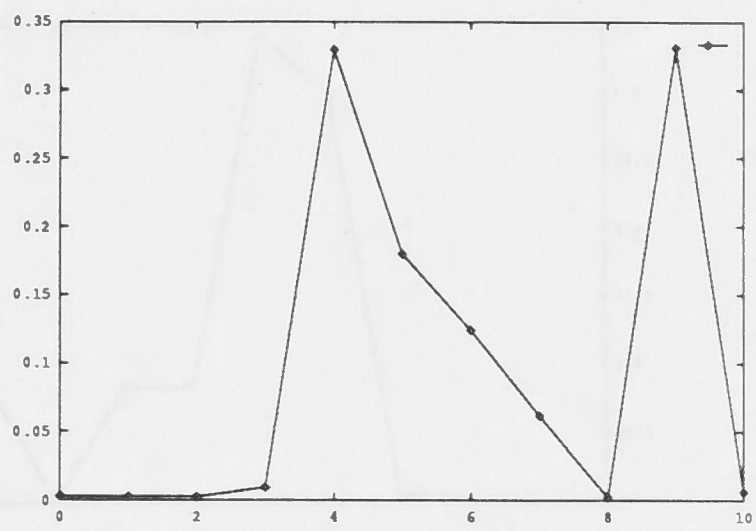
(a)



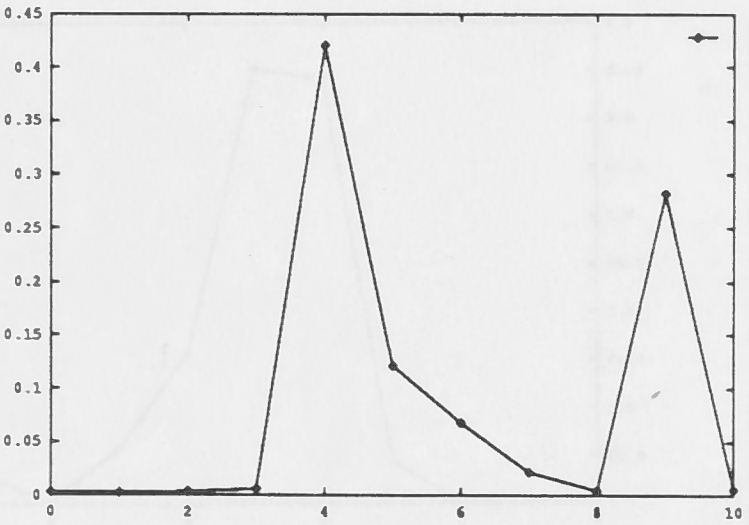
(b)



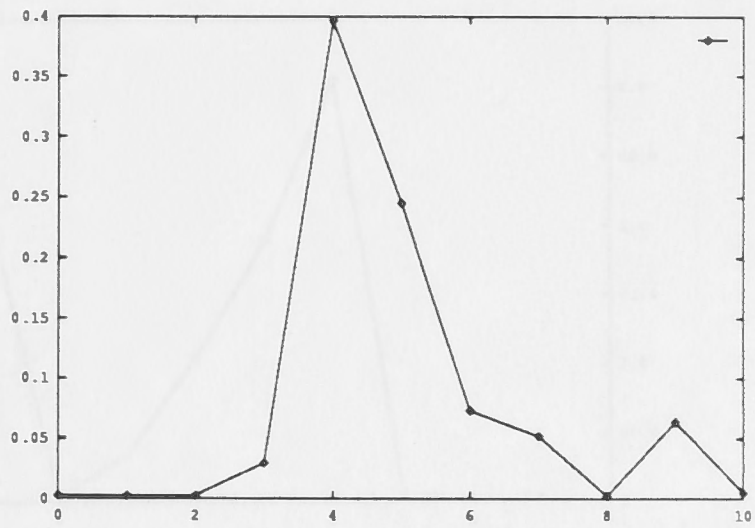
(c)



(d)

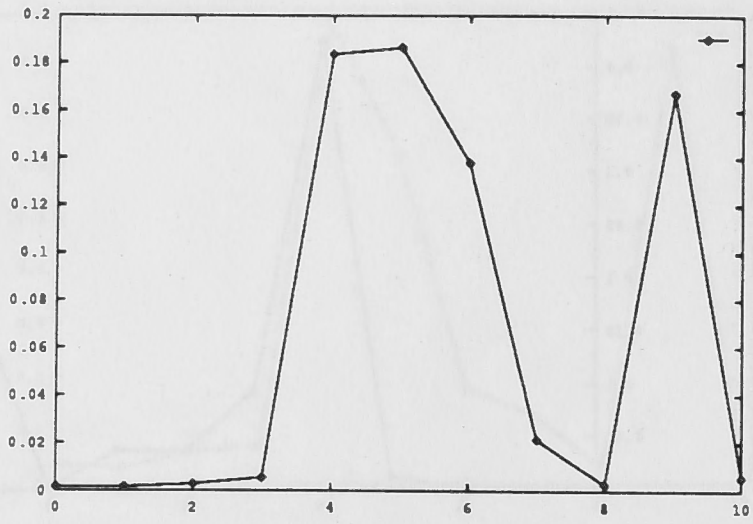


(e)

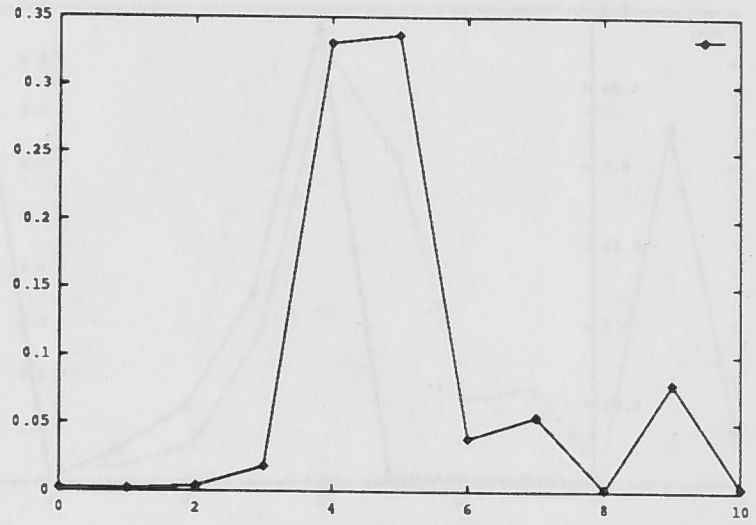


(f)

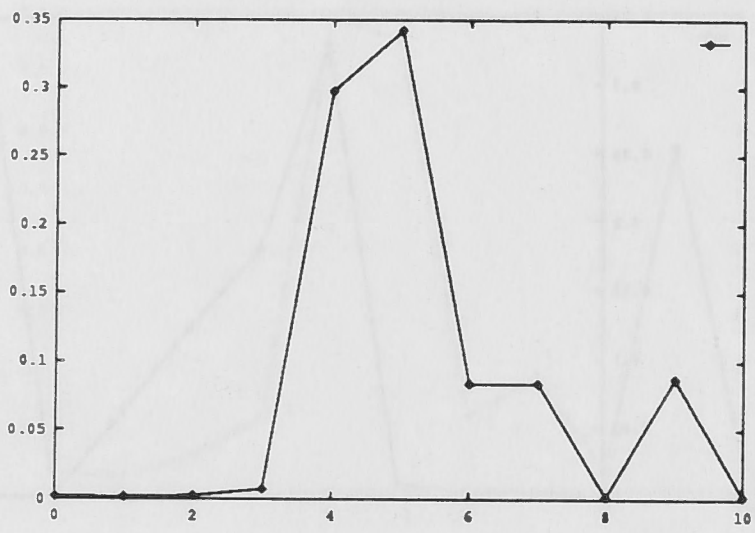
Figure A.20: *Test results of PP's* compact *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



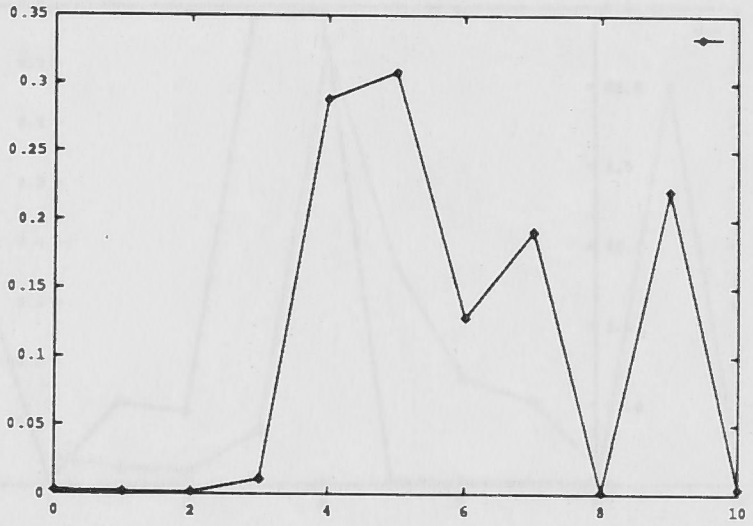
(a)



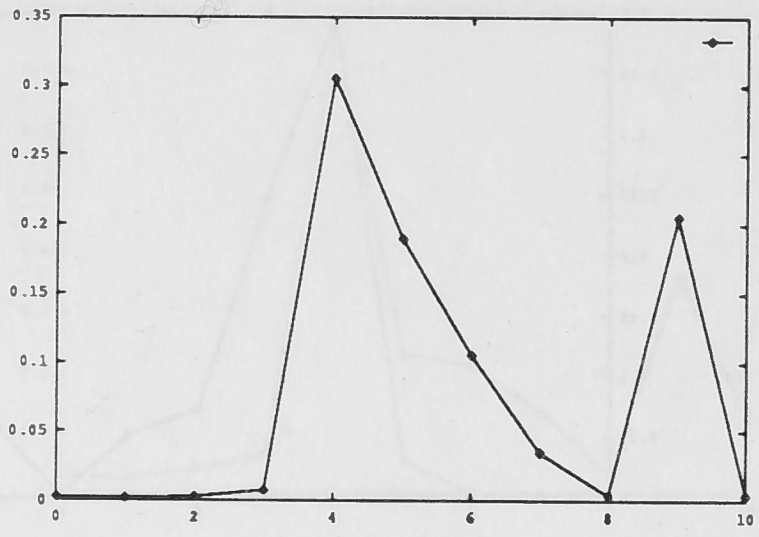
(b)



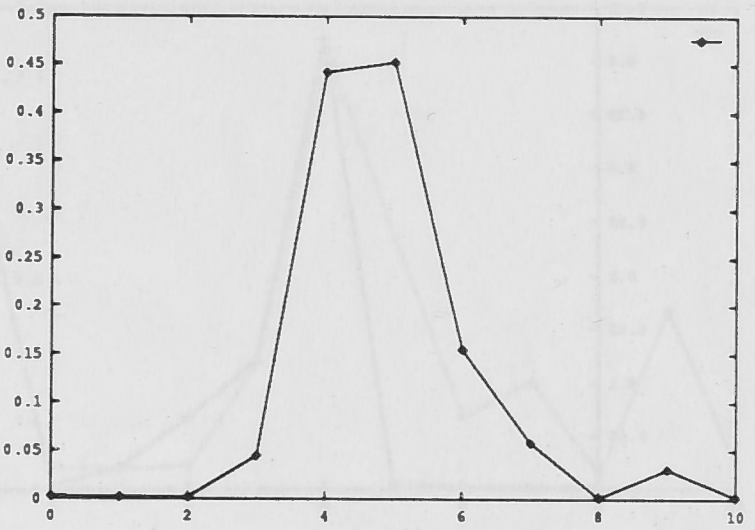
(c)



(d)



(e)

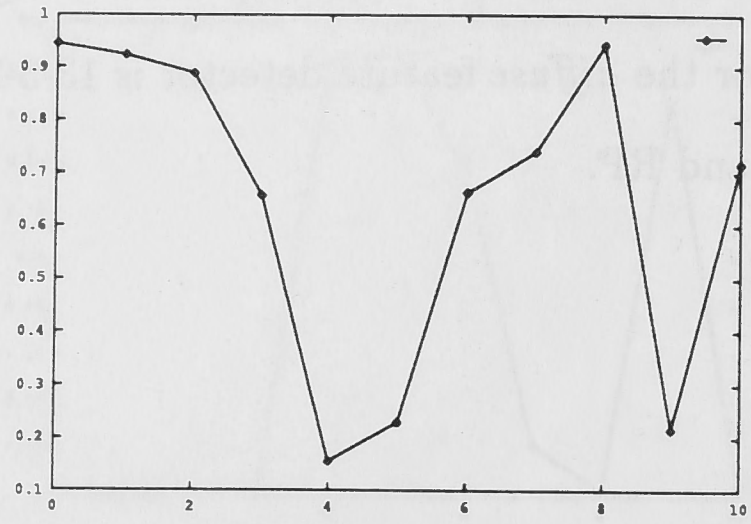


(f)

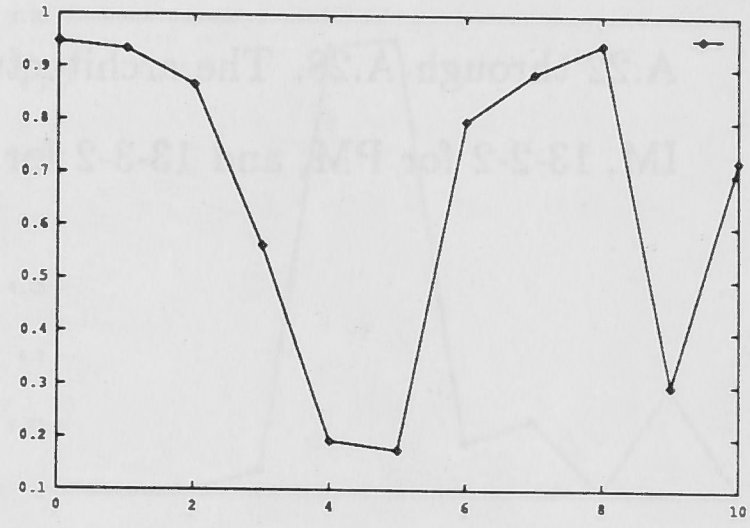
Figure A.21: *Test results of RP's* compact *feature detector tested on PM's 11 pseudo steady-state vowels in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

A.3 Feature "Diffuse"

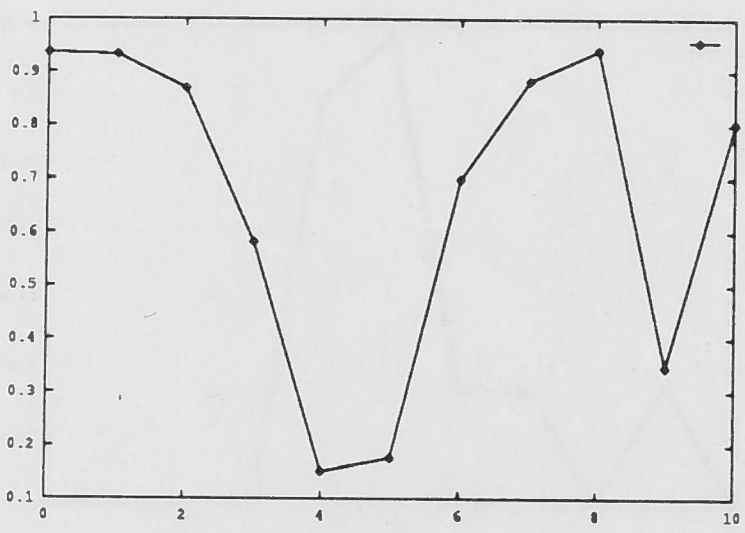
This section presents a complete set of test results of feature *diffuse* detector for each speaker, tested on themselves and on PM respectively, as reported in Figures A.22 through A.28. The architecture for the *diffuse* feature detector is 13-3-2 for IM, 13-2-2 for PM, and 13-3-2 for PP and RP.



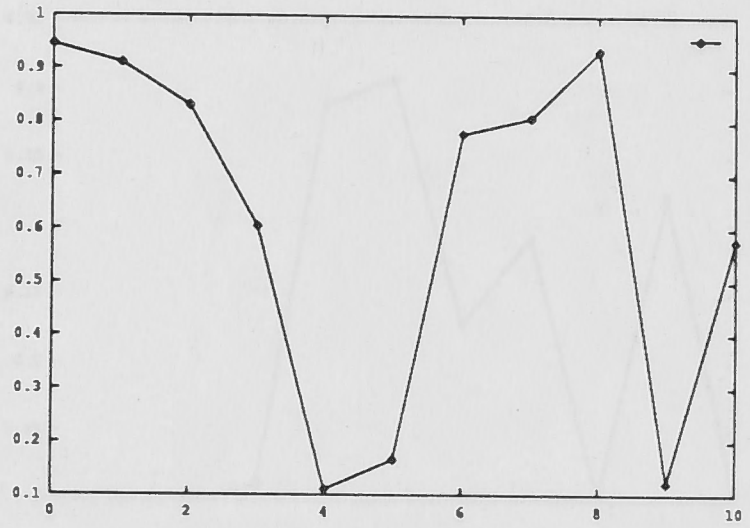
(a)



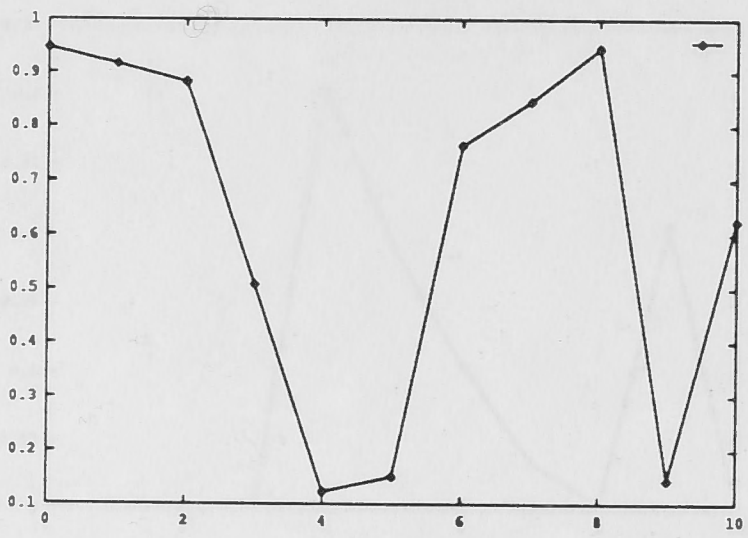
(b)



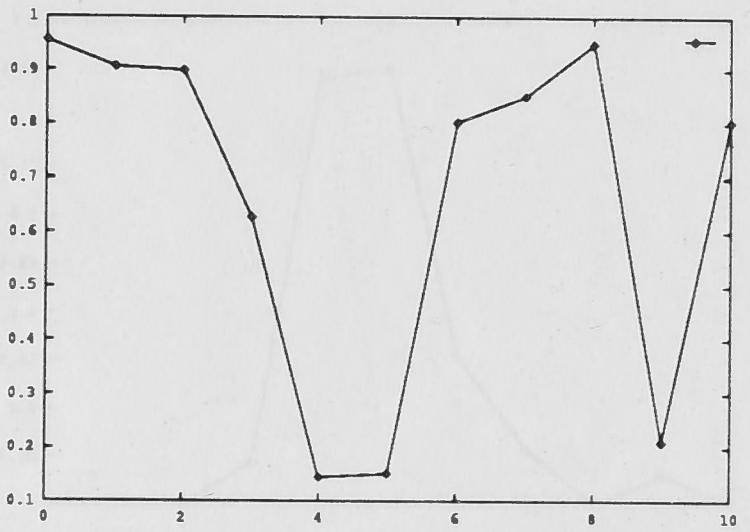
(c)



(d)

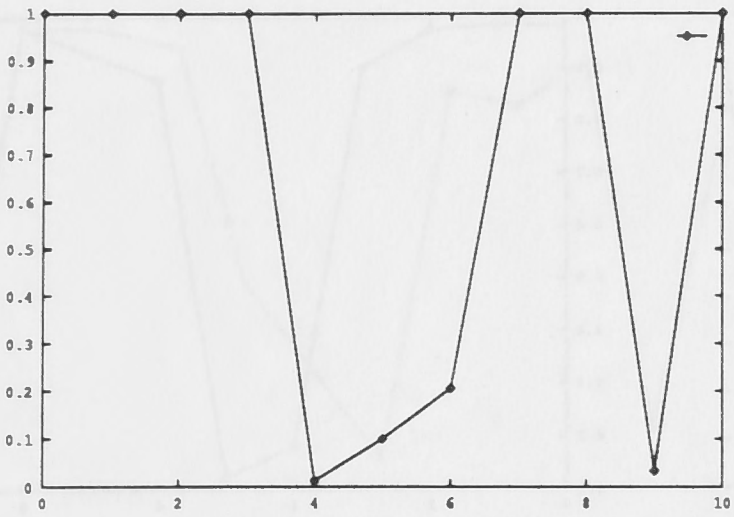


(e)

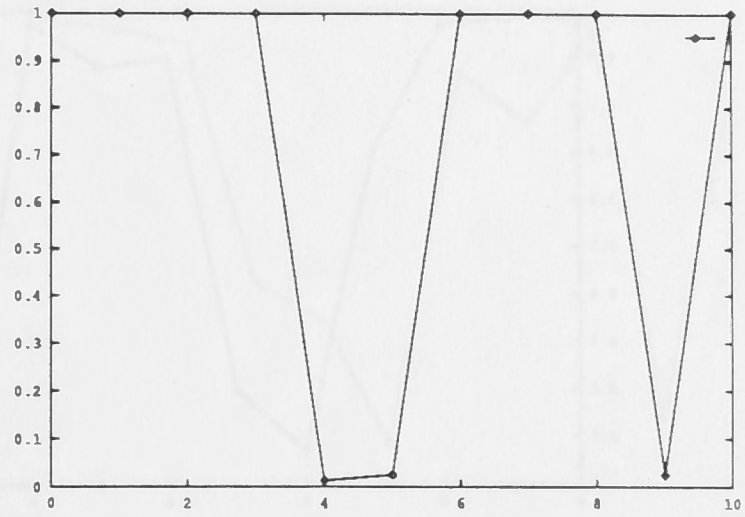


(f)

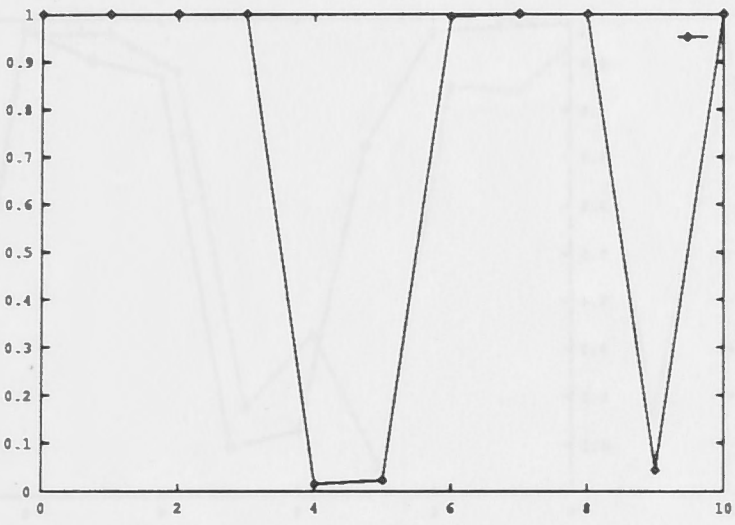
Figure A.22: *Test results of IM's* diffuse *feature detector tested on IM's 11 pseudo steady-state vowel in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



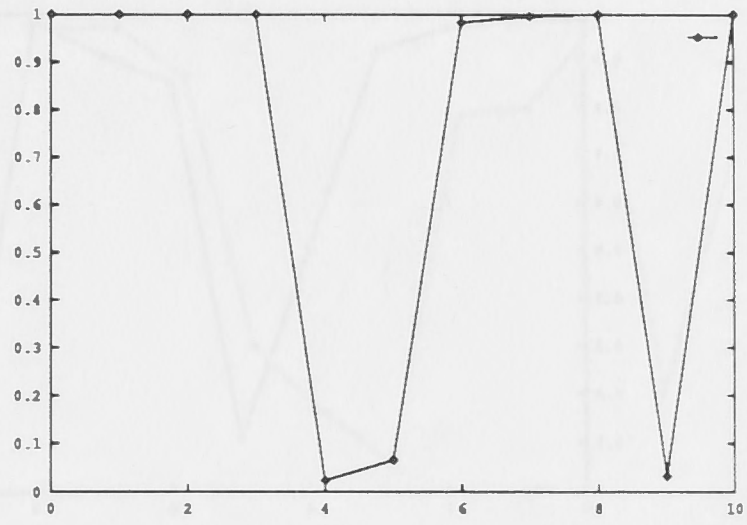
(a)



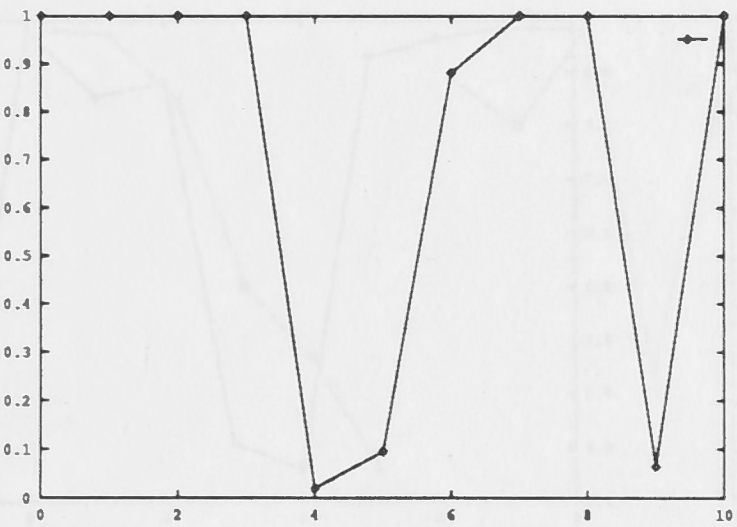
(b)



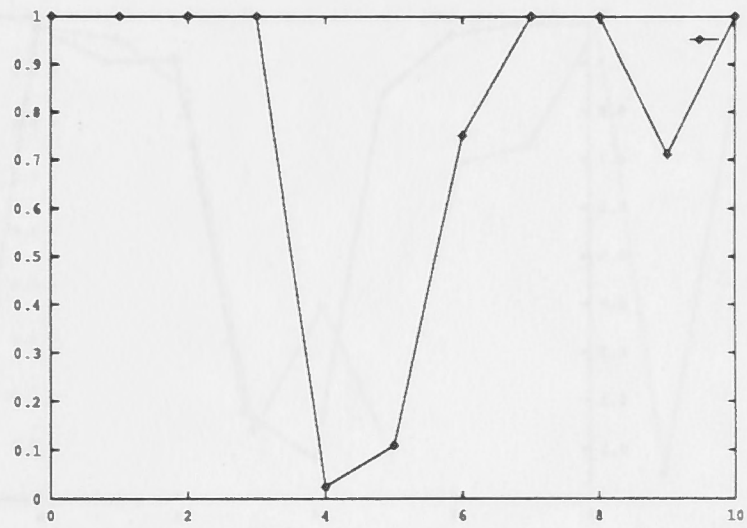
(c)



(d)

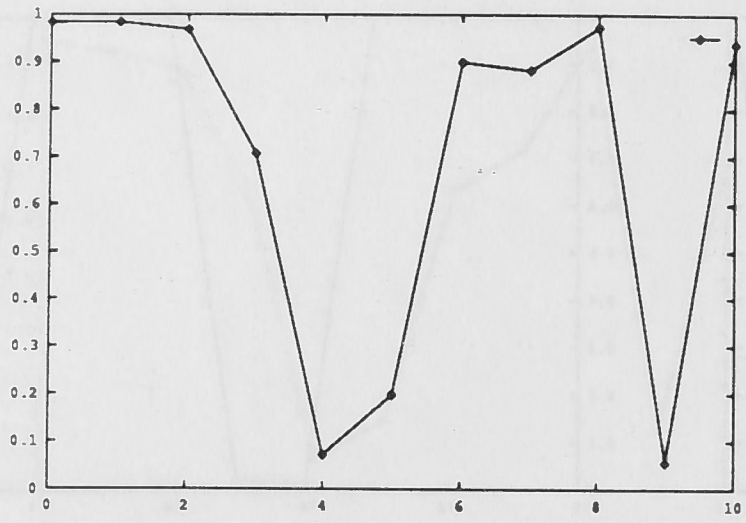


(e)

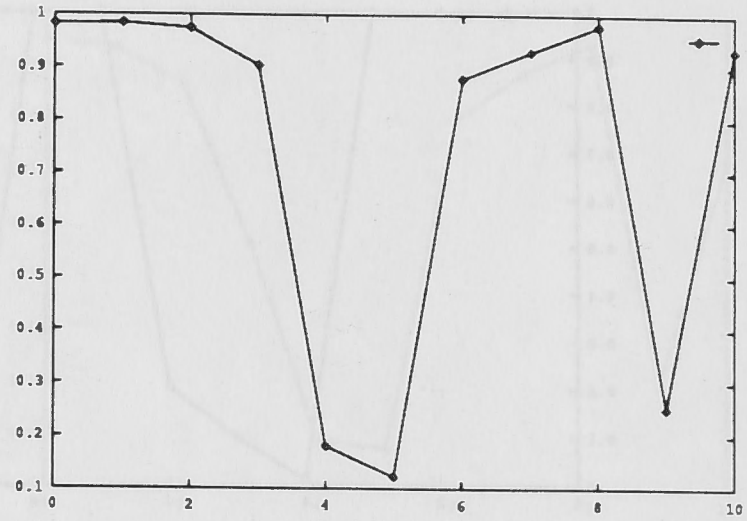


(f)

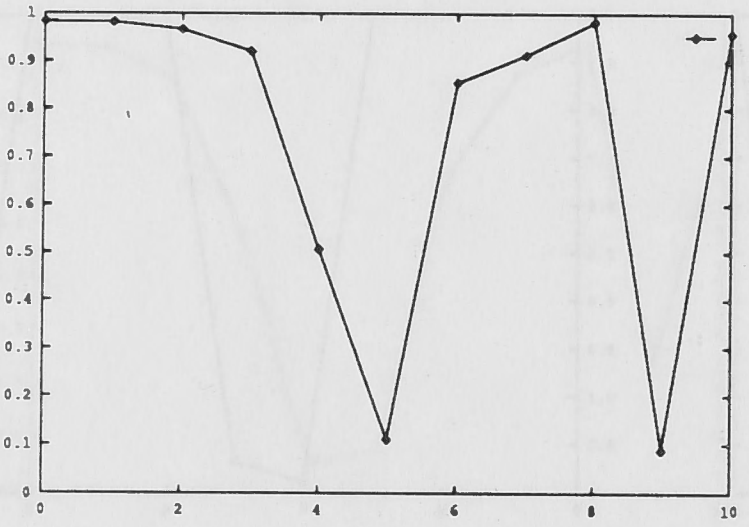
Figure A.23: *Test results of PM's* diffuse *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



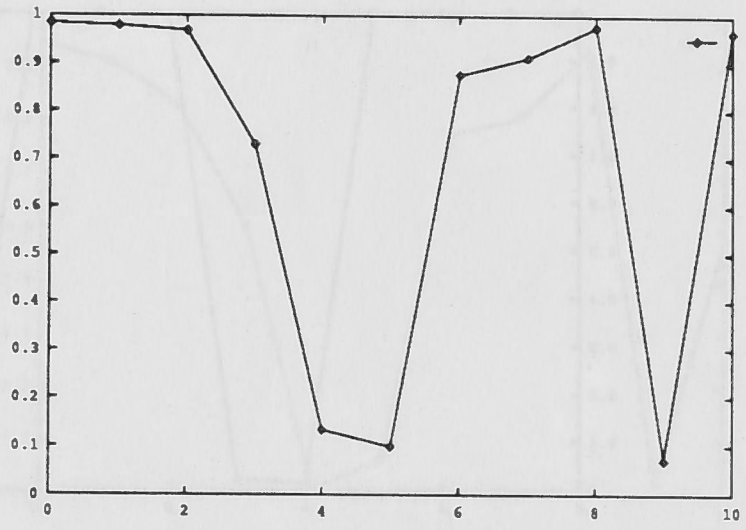
(a)



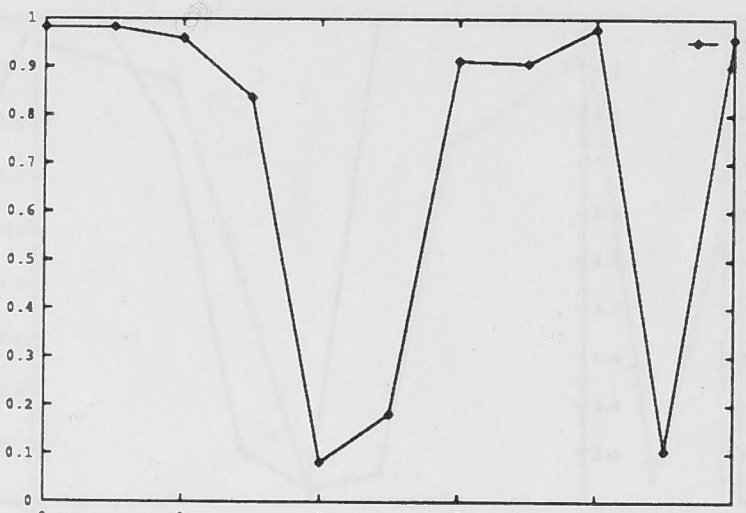
(b)



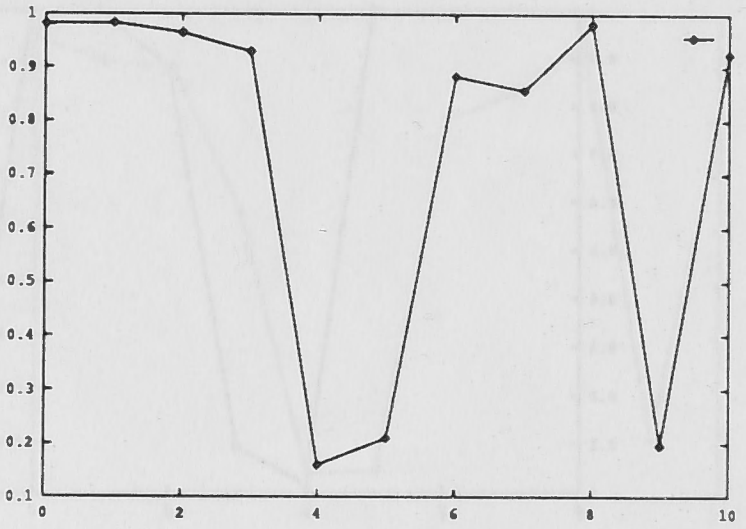
(c)



(d)



(e)



(f)

Figure A.24: *Test results of PP's* diffuse *feature detector tested on PP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

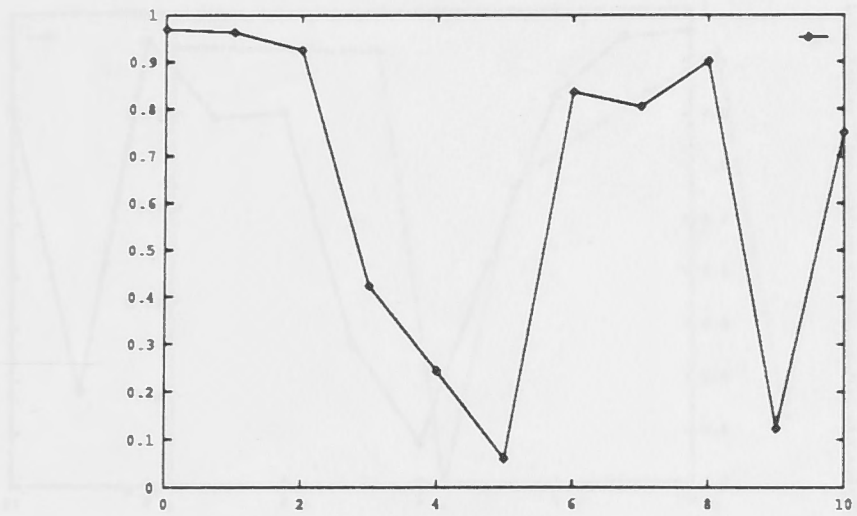
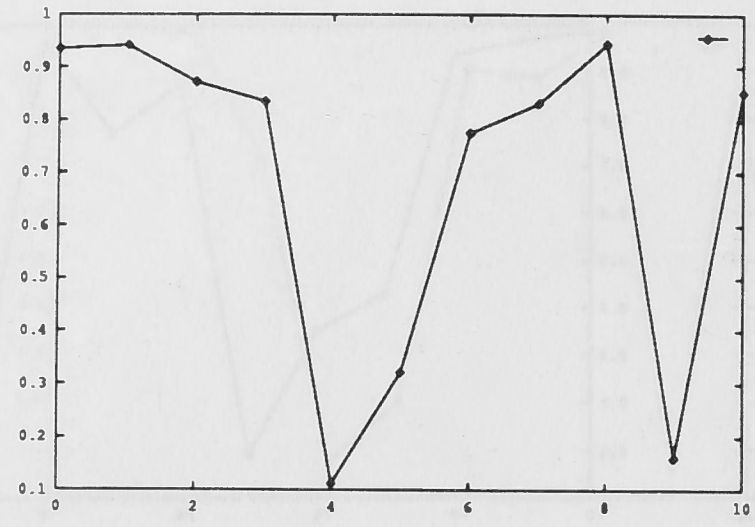
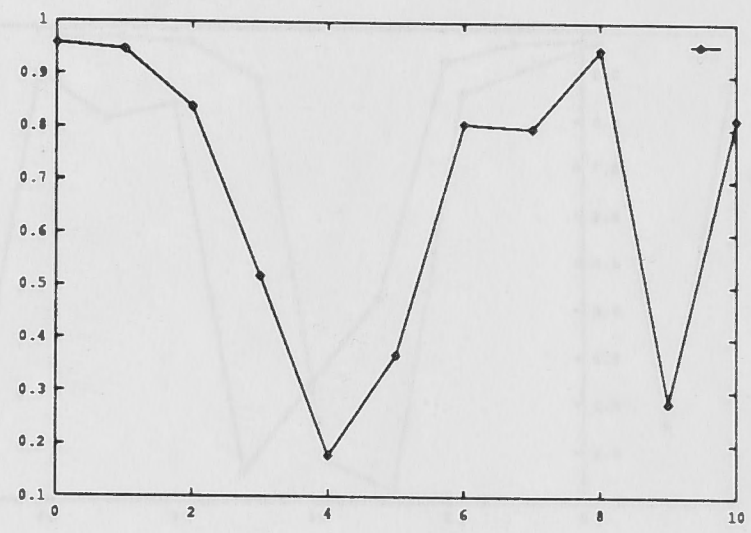


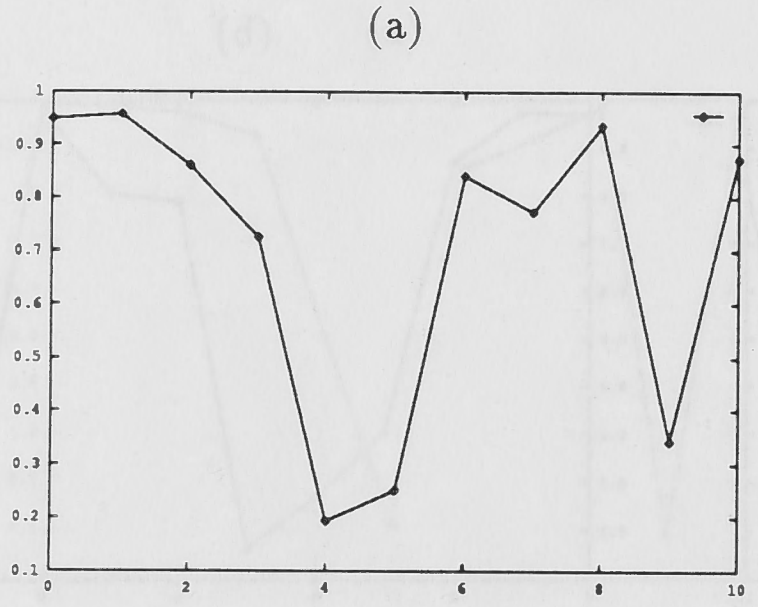
Figure A.25: *Test results of RP's* diffuse *feature detector tested on RP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



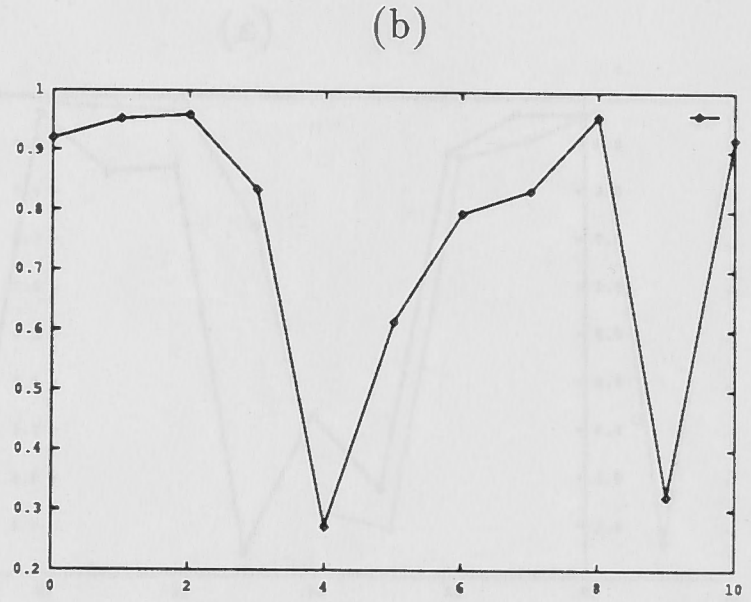
(a)



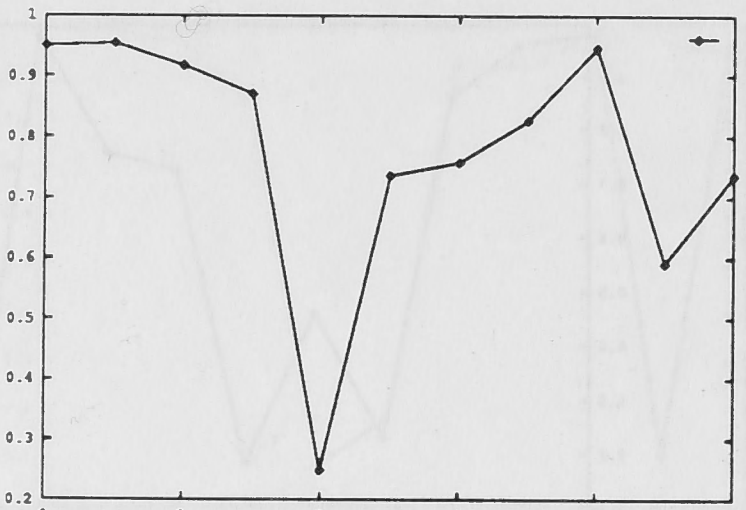
(b)



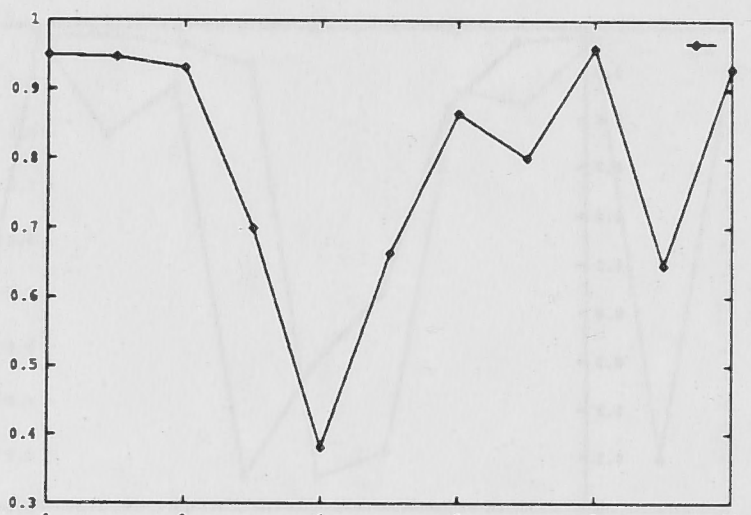
(c)



(d)



(e)



(f)

Figure A.26: *Test results of IM's* diffuse *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

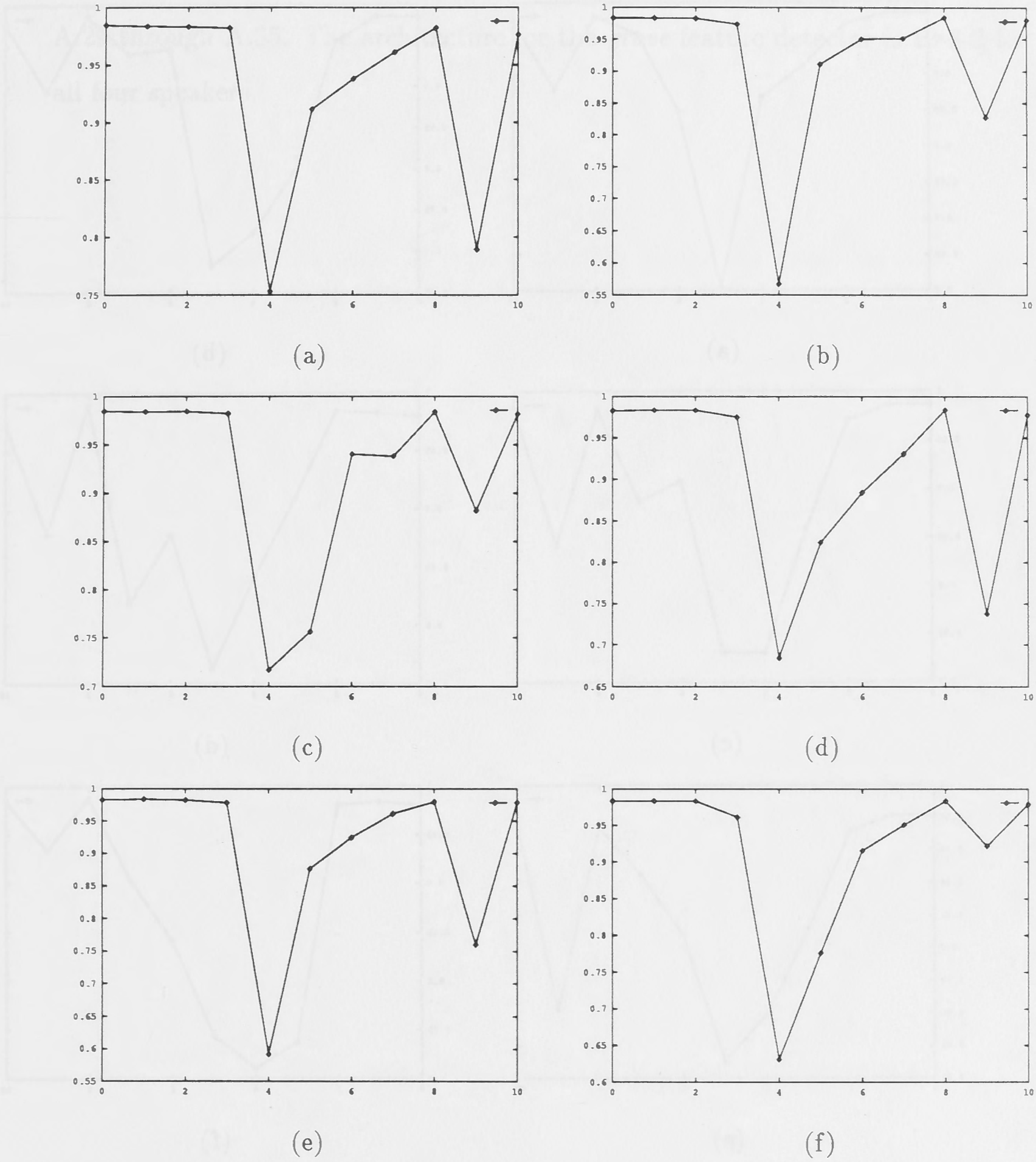
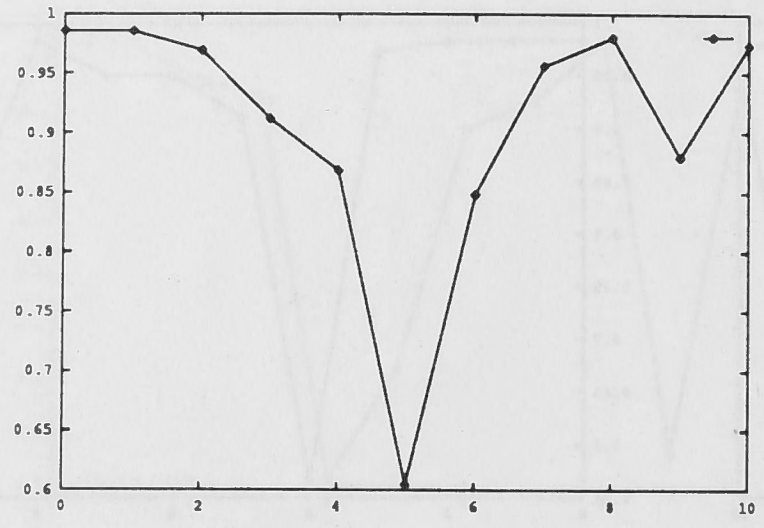
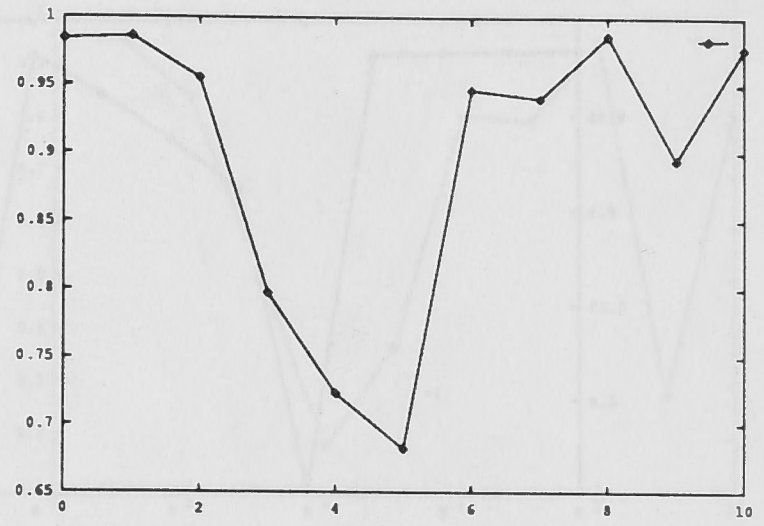


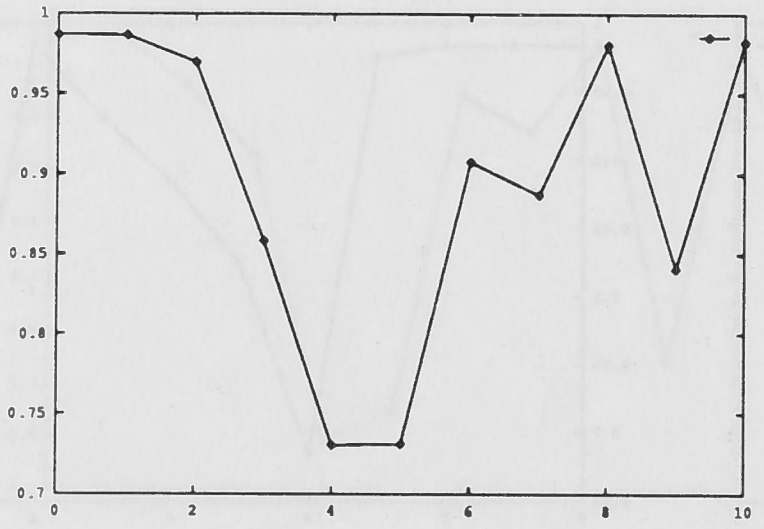
Figure A.27: *Test results of PP's* diffuse *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



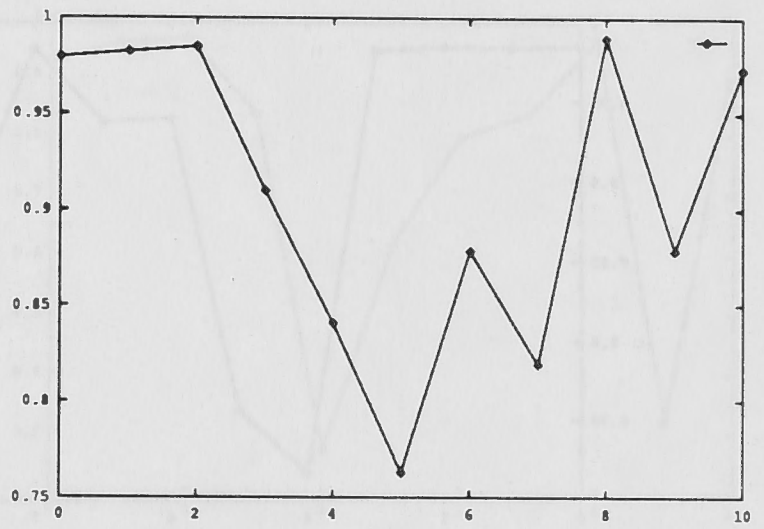
(a)



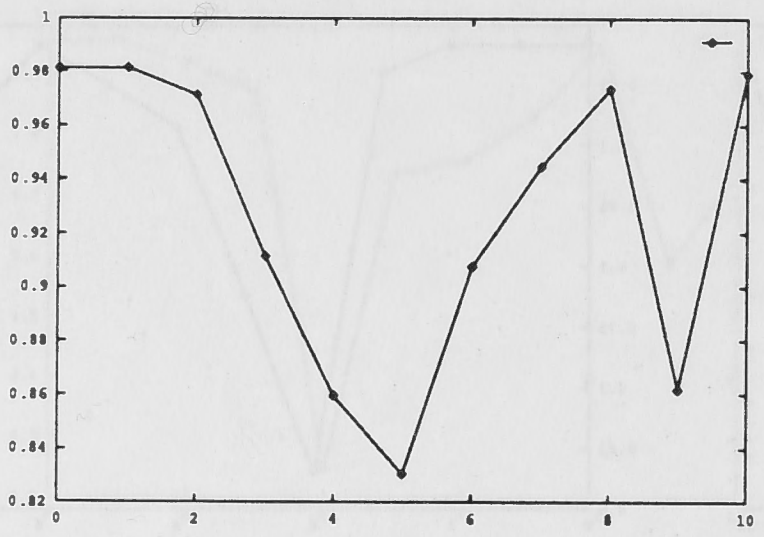
(b)



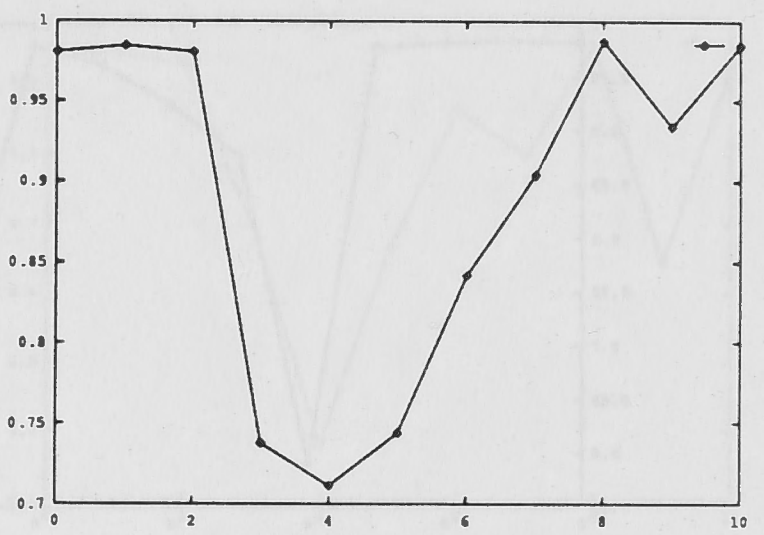
(c)



(d)



(e)

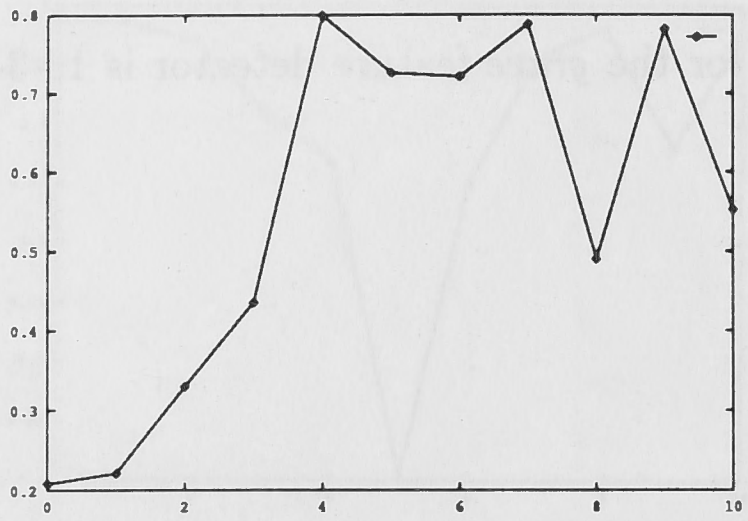


(f)

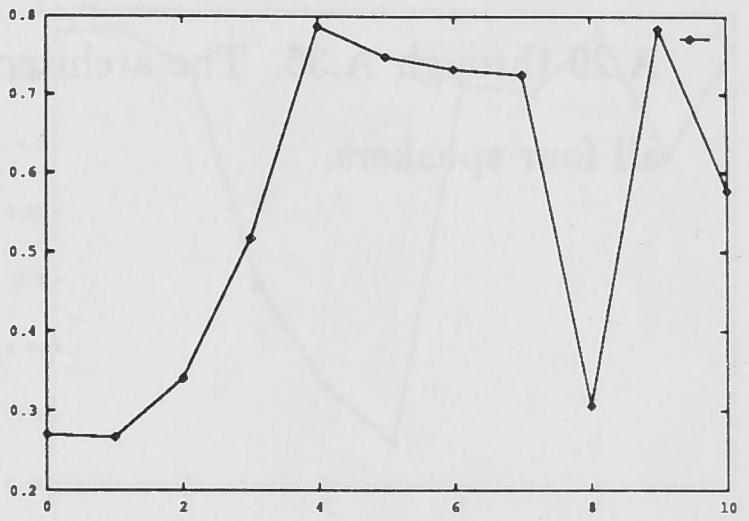
Figure A.28: *Test results of RP's* diffuse *feature detector tested on PM's 11 pseudo static state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

A.4 Feature "Grave"

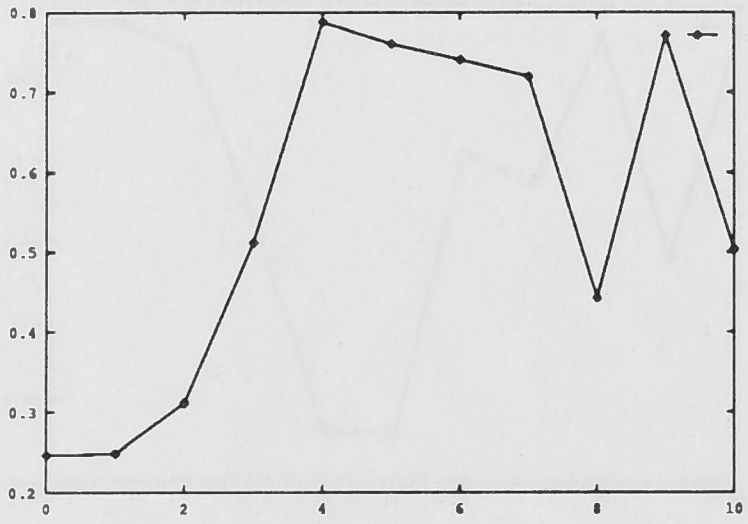
This section presents a complete set of test results of feature *grave* detector for each speaker tested on themselves and on PM respectively, as reported in Figures A.29 through A.35. The architecture for the *grave* feature detector is 13-3-2 for all four speakers.



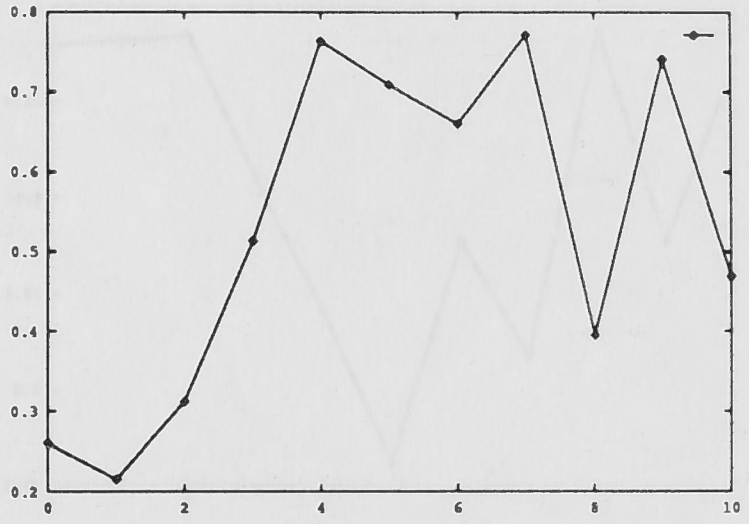
(a)



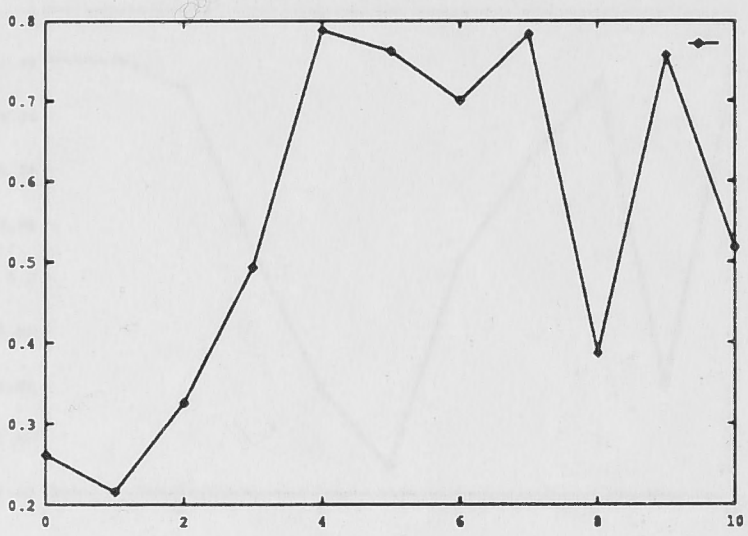
(b)



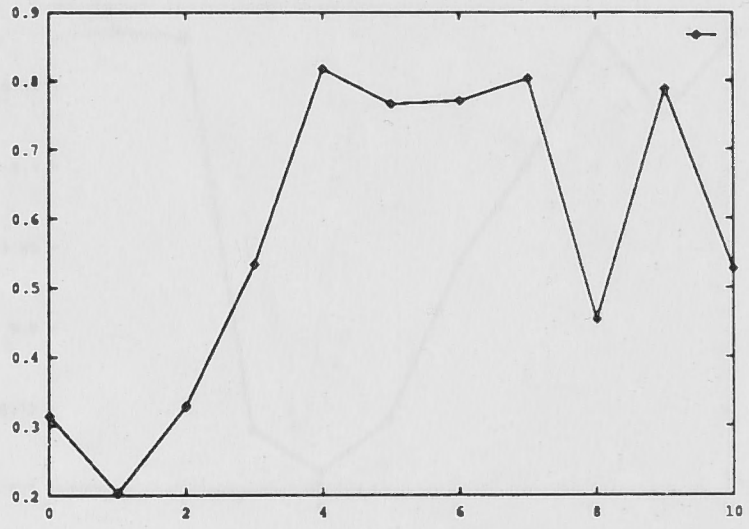
(c)



(d)

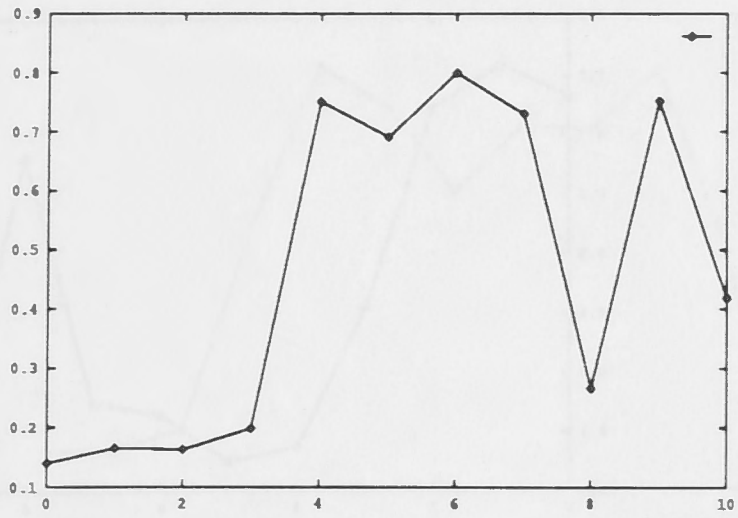


(e)

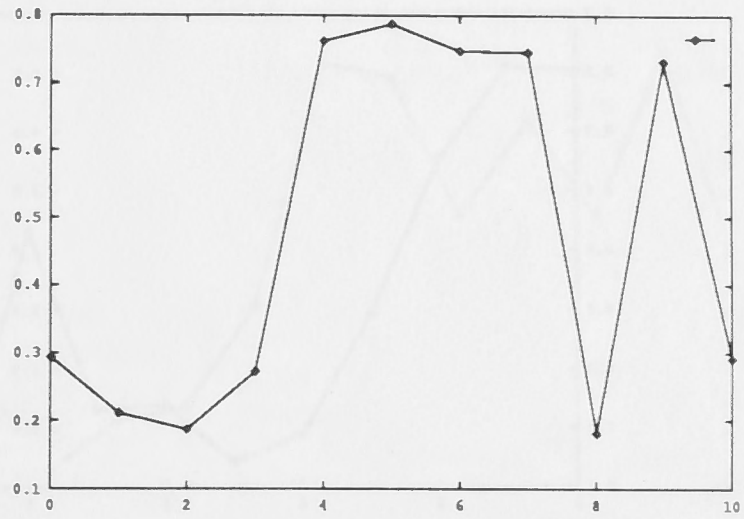


(f)

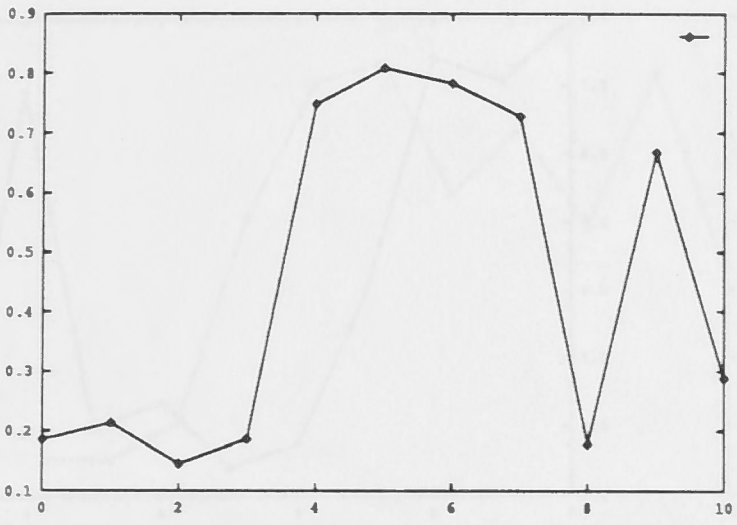
Figure A.29: *Test results of IM's* grave *feature detector tested on IM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



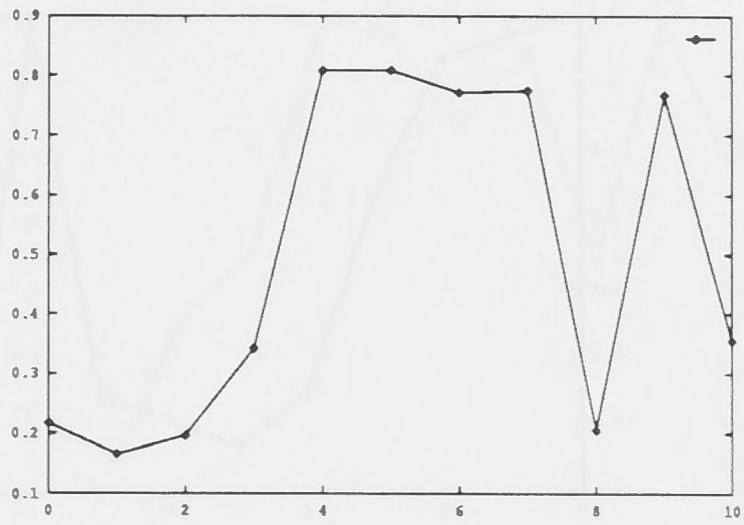
(a)



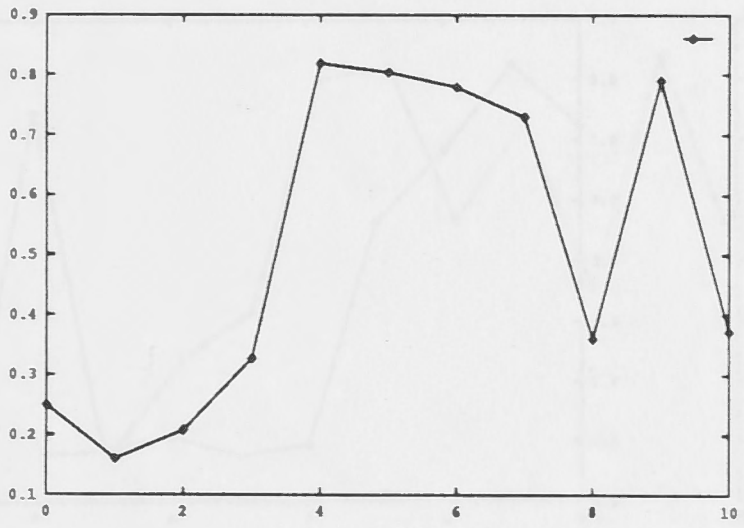
(b)



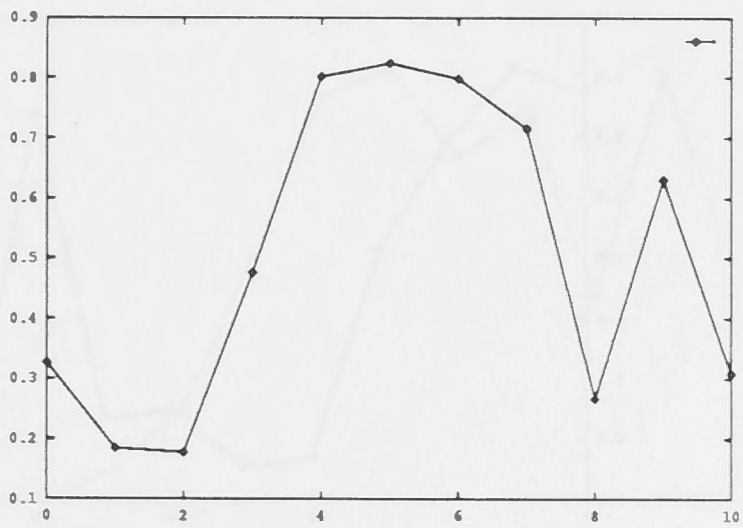
(c)



(d)

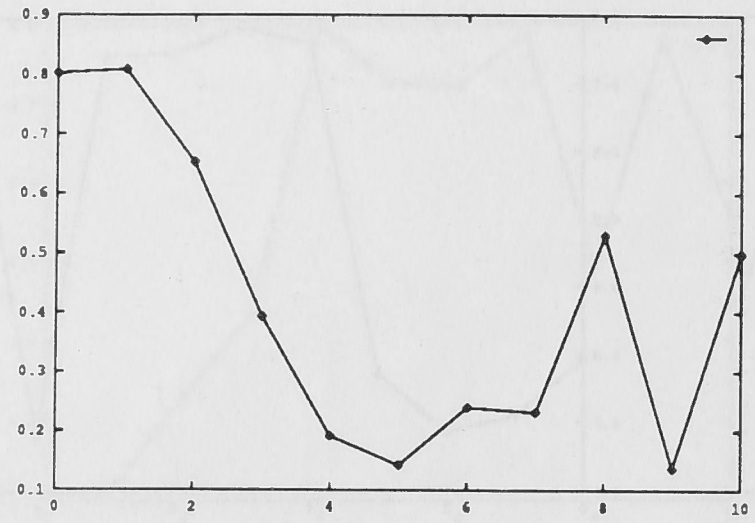


(e)

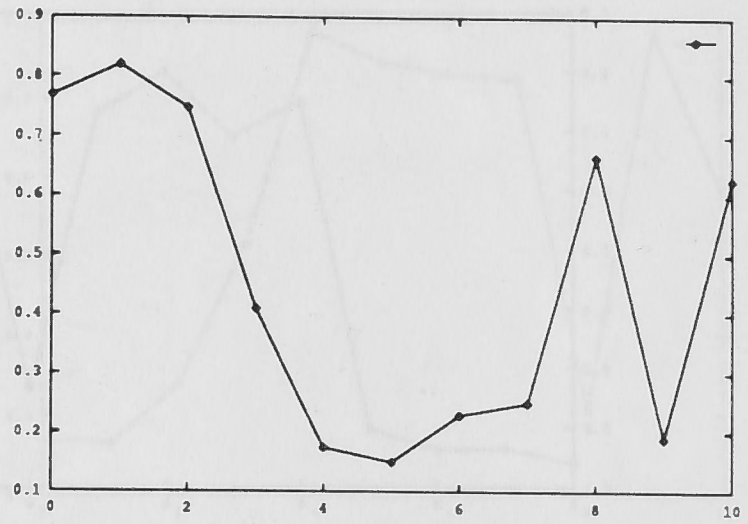


(f)

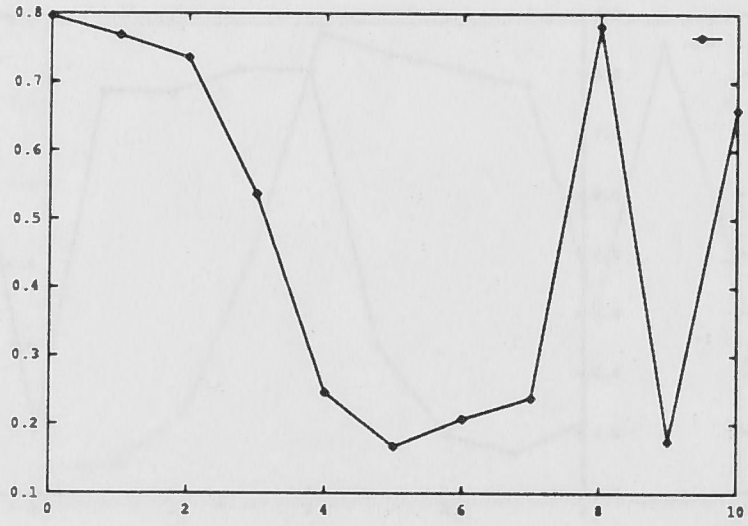
Figure A.30: *Test results of PM's* grave *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



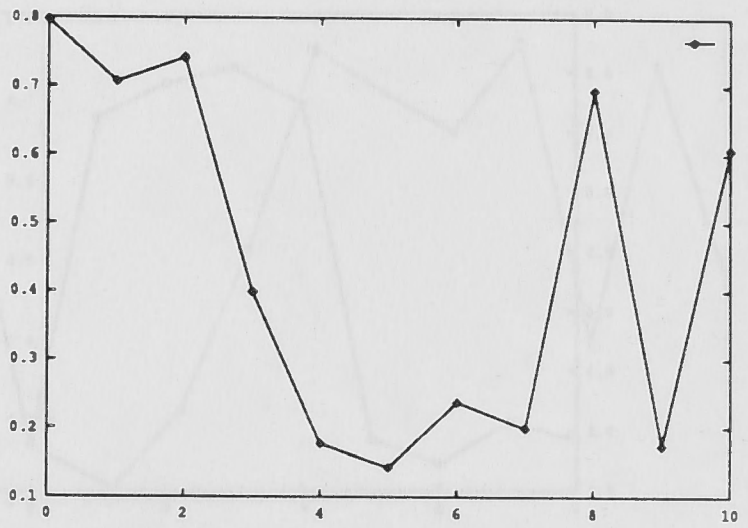
(a)



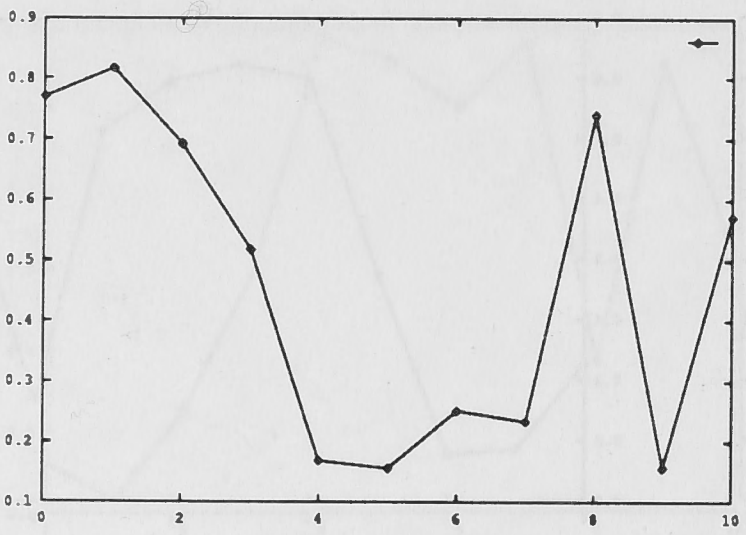
(b)



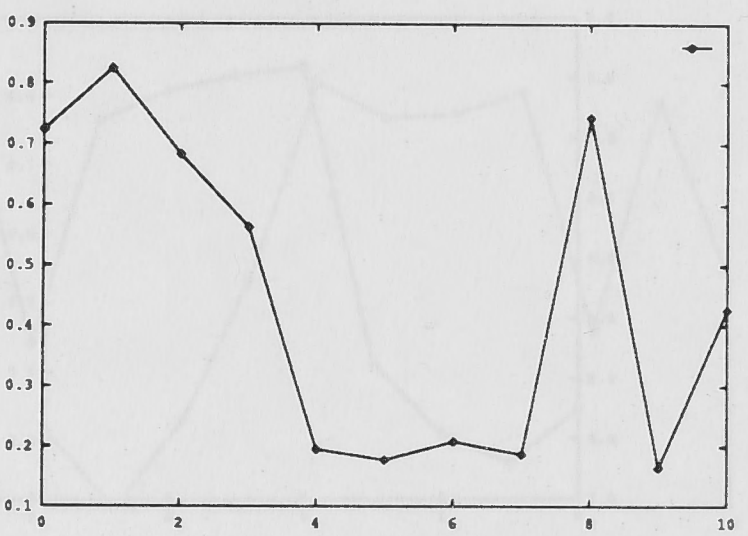
(c)



(d)

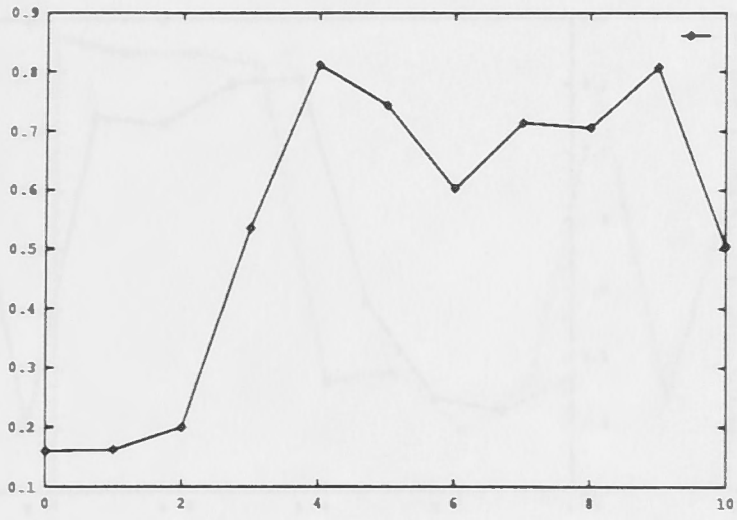


(e)

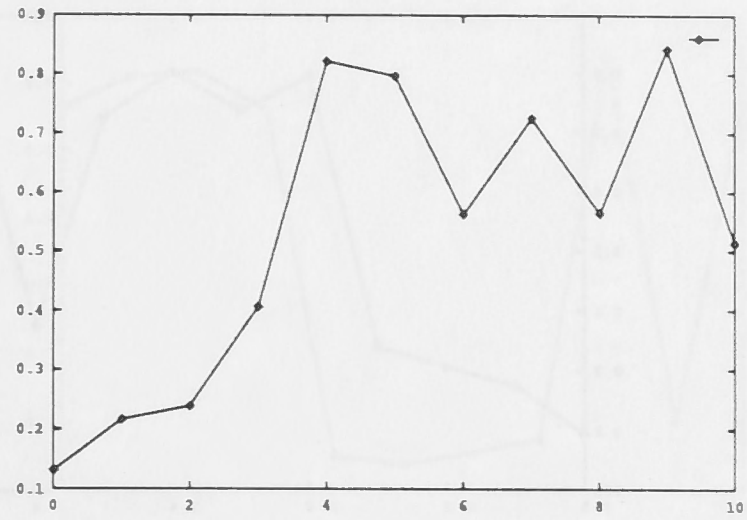


(f)

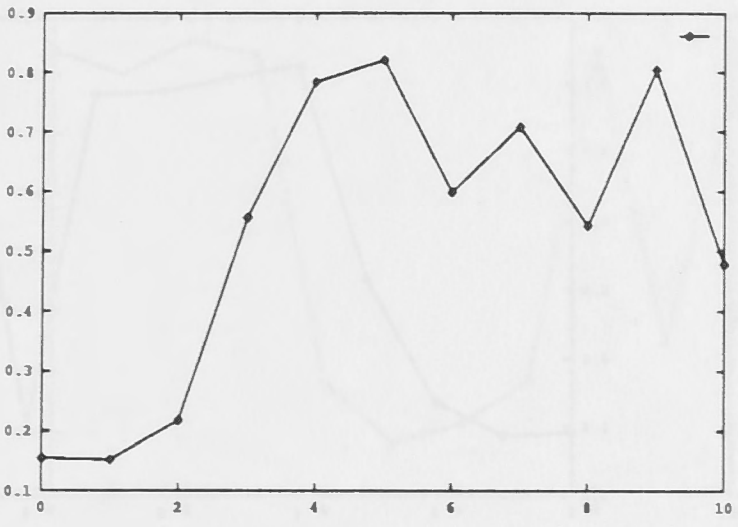
Figure A.31: *Test results of PP's* grave *feature detector tested on PP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



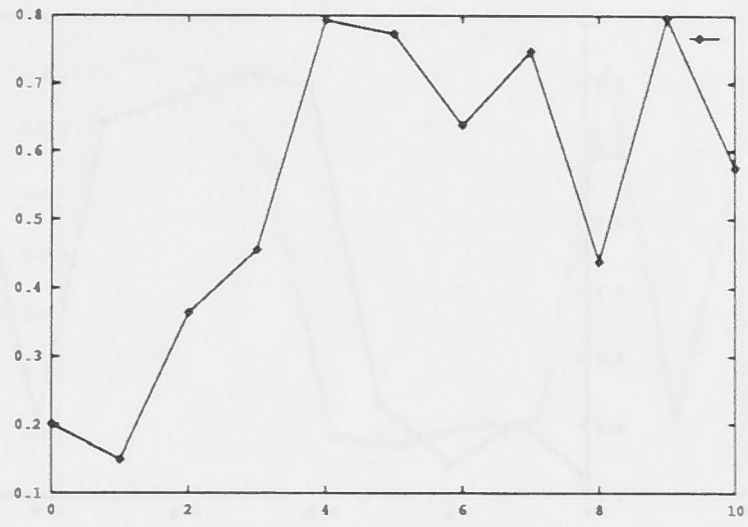
(a)



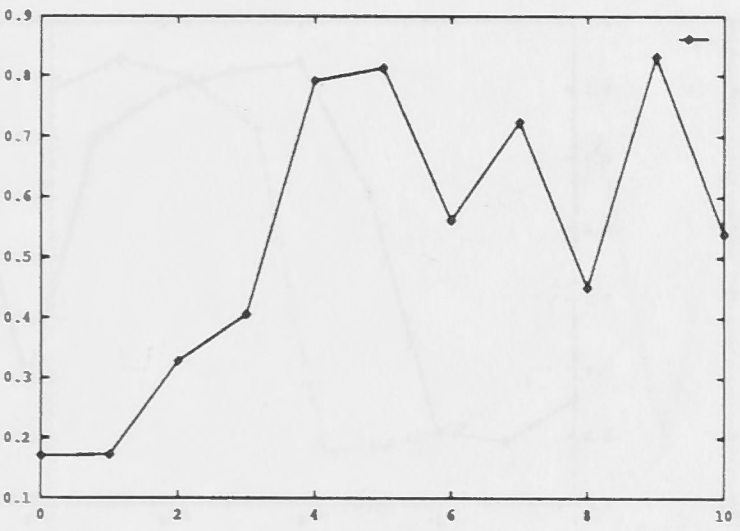
(b)



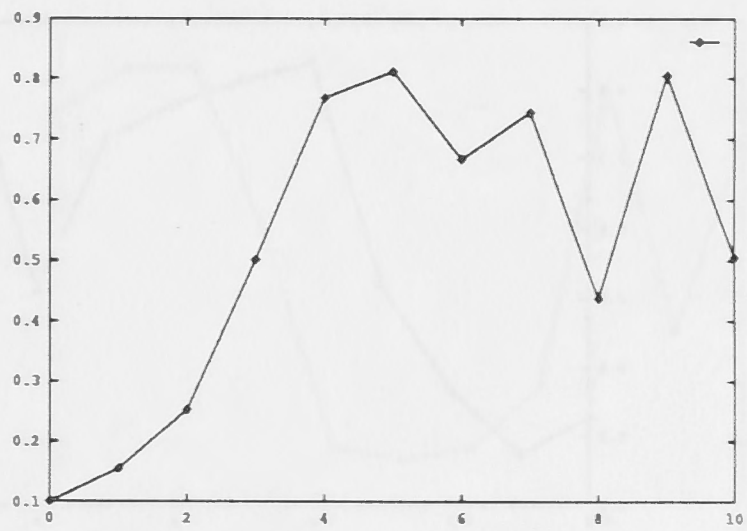
(c)



(d)

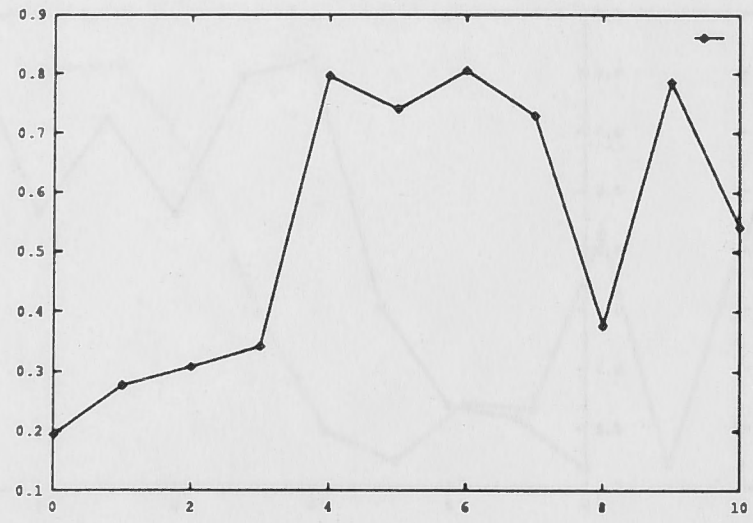


(e)

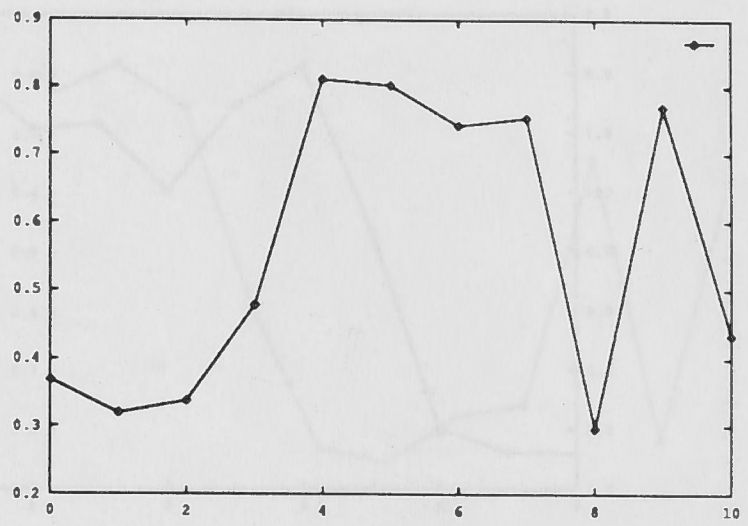


(f)

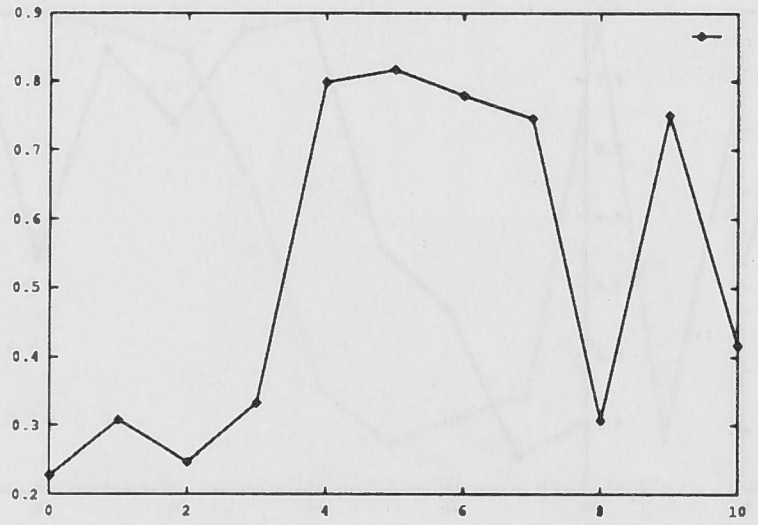
Figure A.32: *Test results of RP's* grave *feature detector tested on RP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



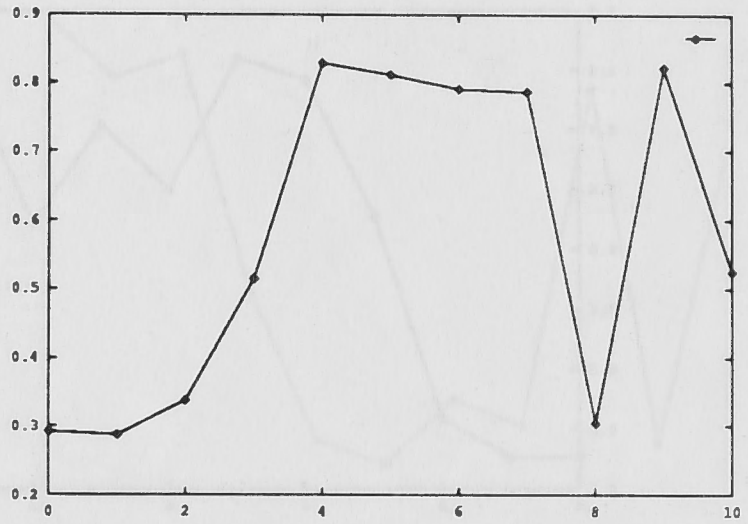
(a)



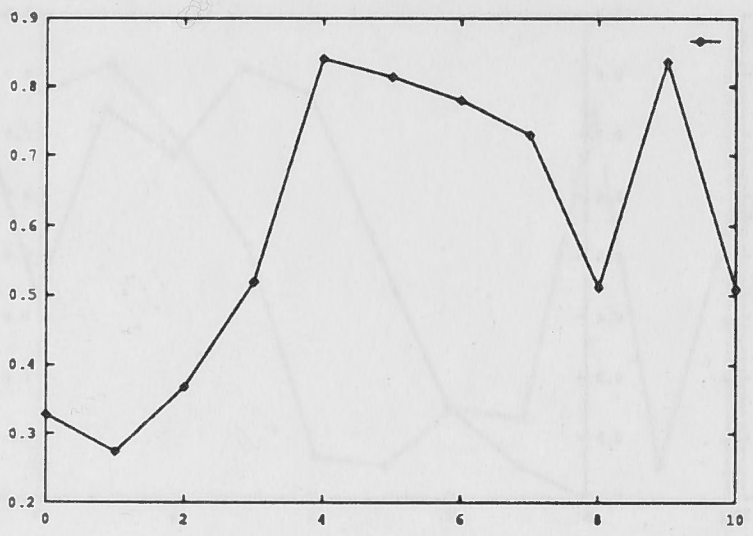
(b)



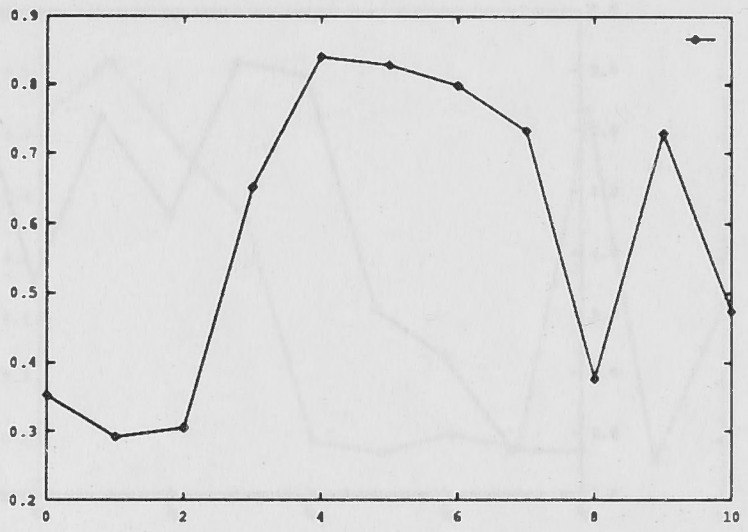
(c)



(d)



(e)



(f)

Figure A.33: *Test results of IM's* grave *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

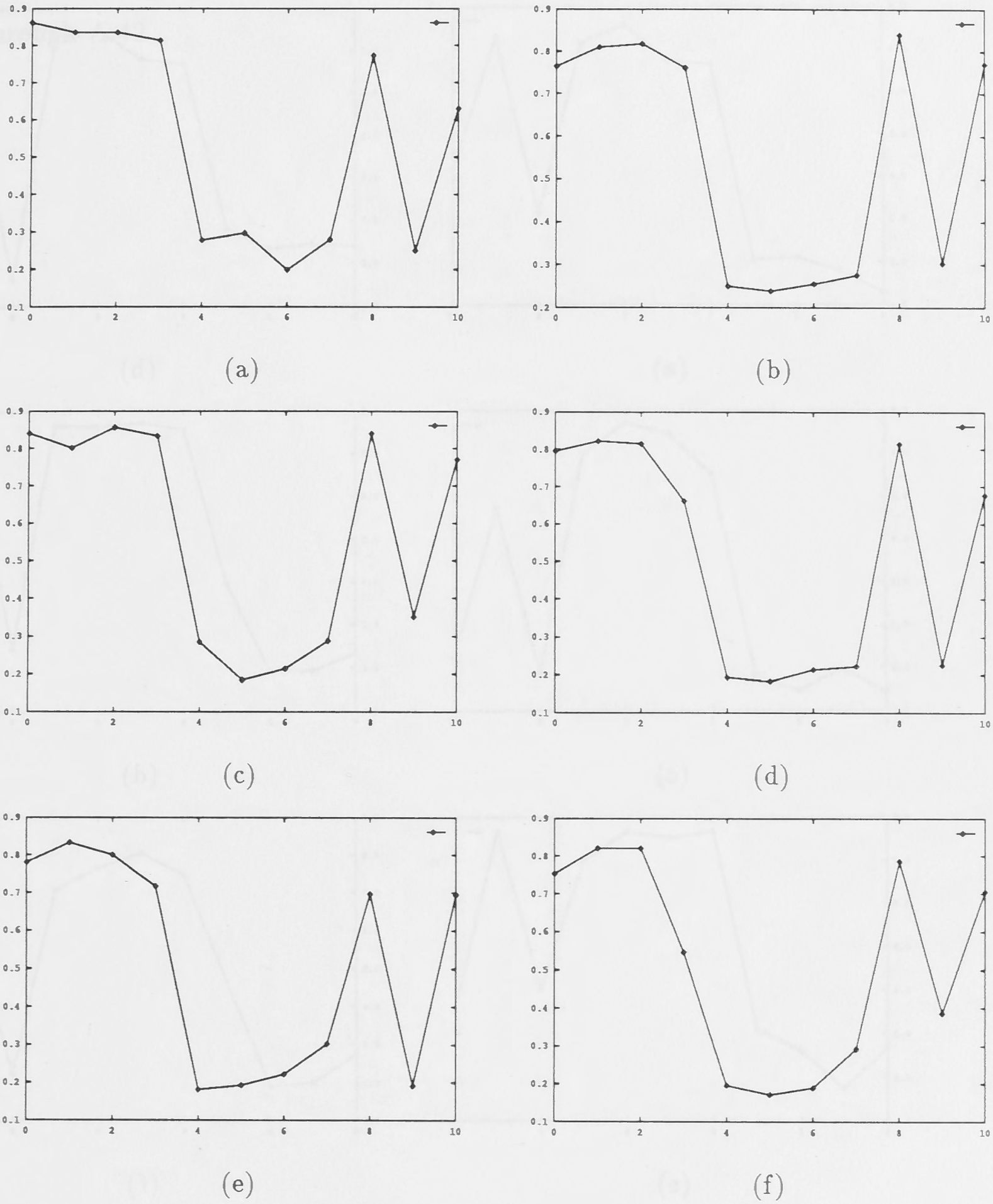
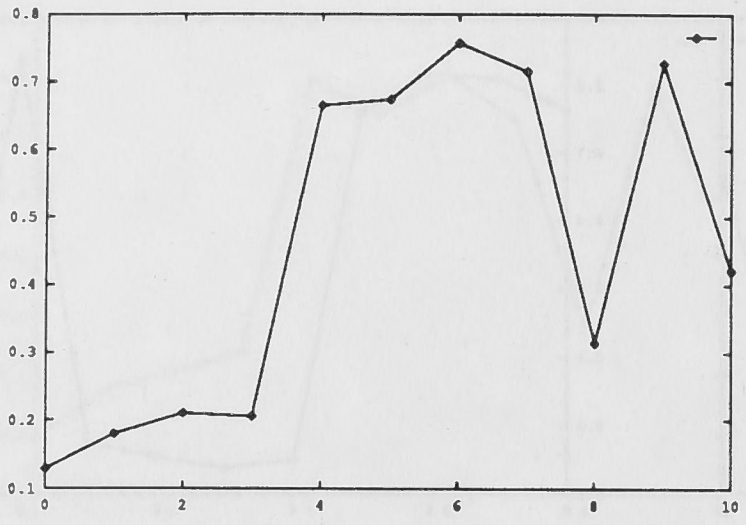
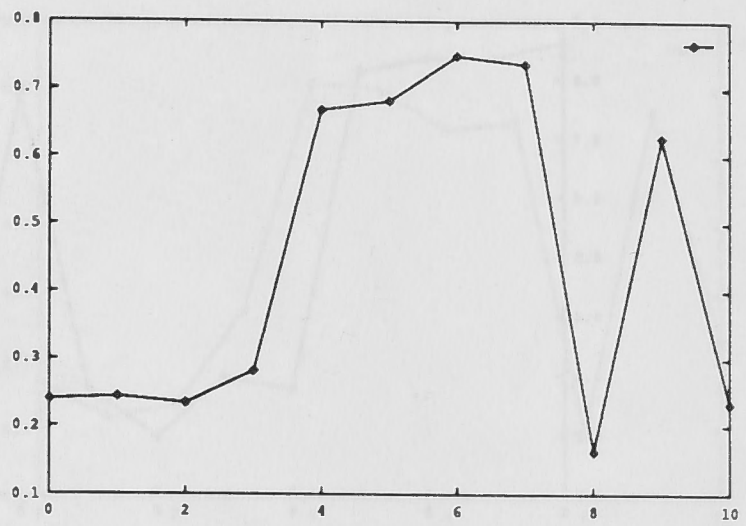


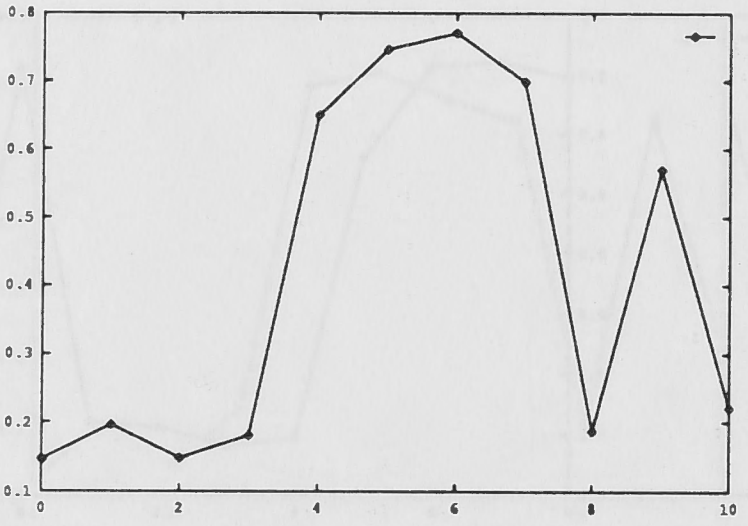
Figure A.34: *Test results of PP's* grave *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



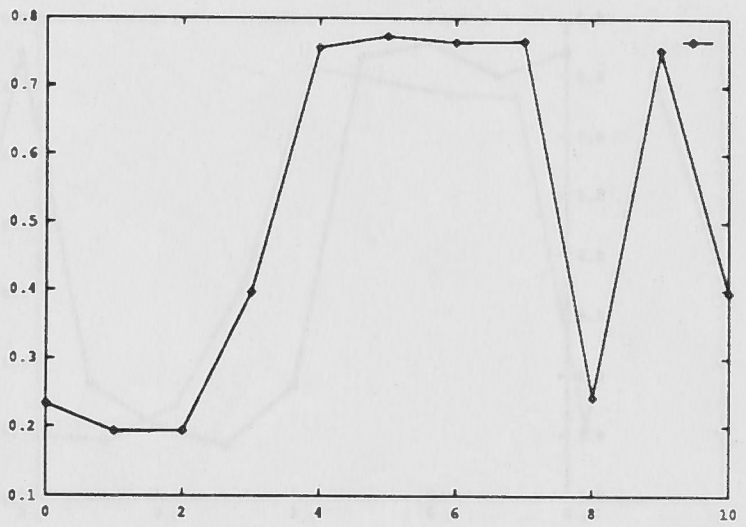
(a)



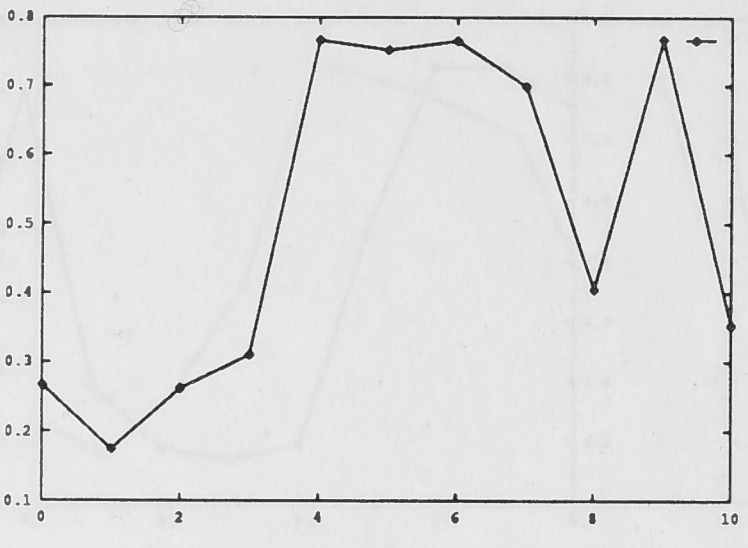
(b)



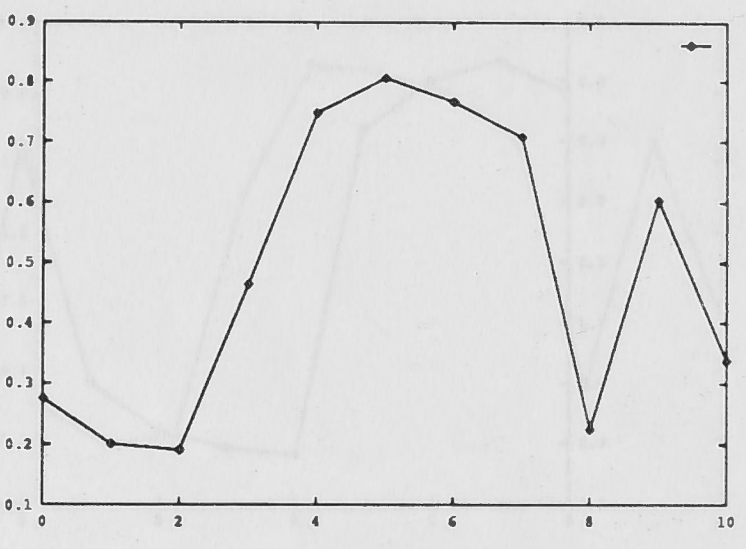
(c)



(d)



(e)

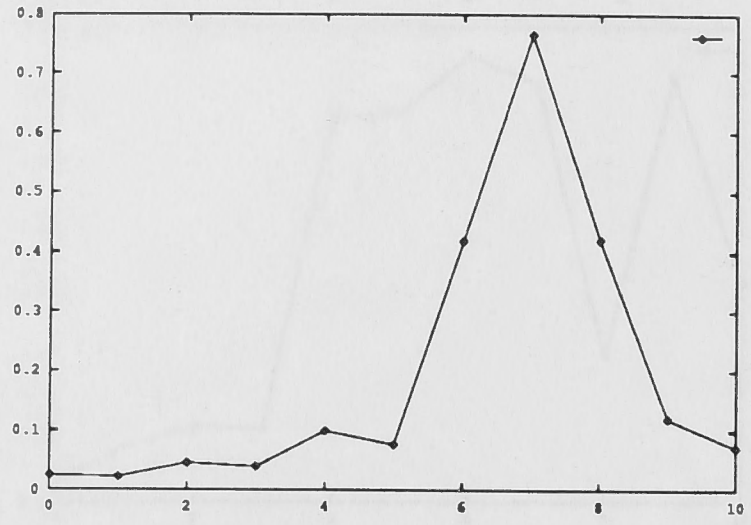


(f)

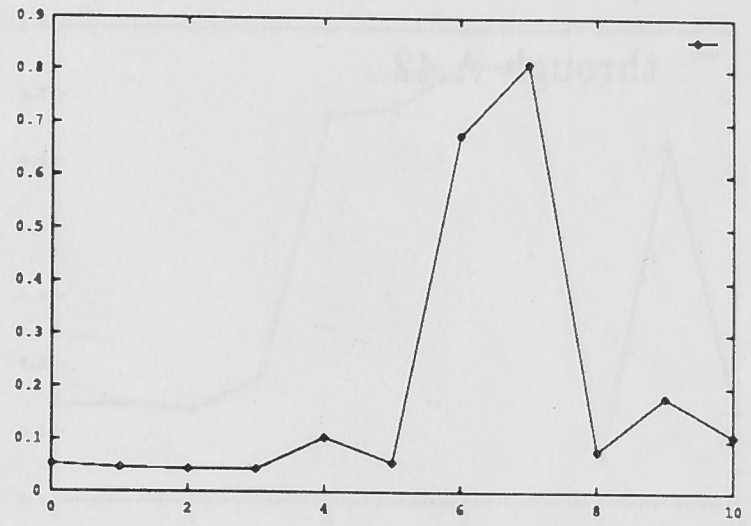
Figure A.35: *Test results of RP's* grave *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

A.5 Feature "Flat"

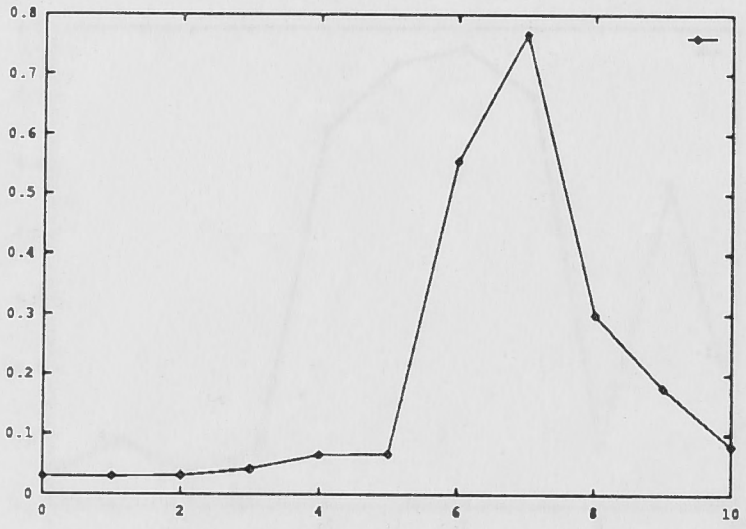
This section presents a complete set of test results of feature *flat* detector for each speaker tested on themselves and on PM respectively, as reported in Figures A.36 through A.42.



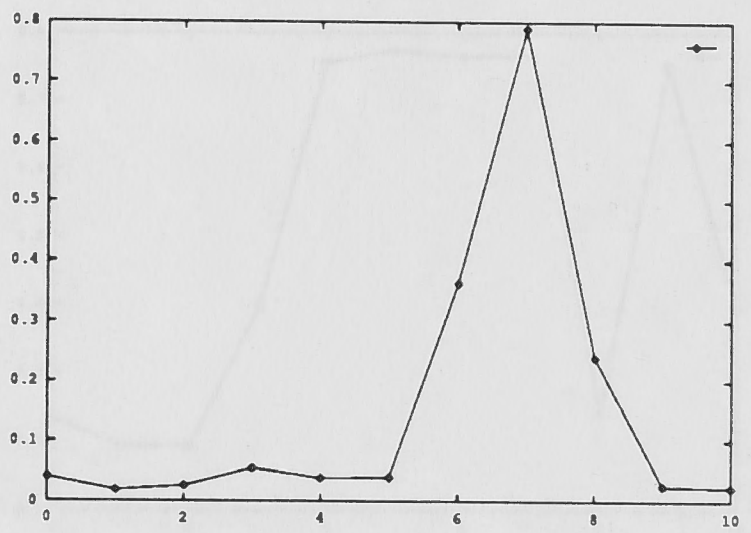
(a)



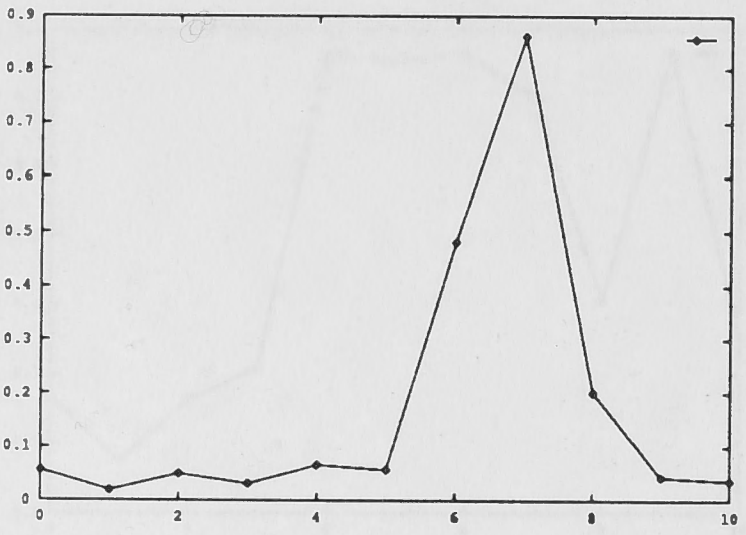
(b)



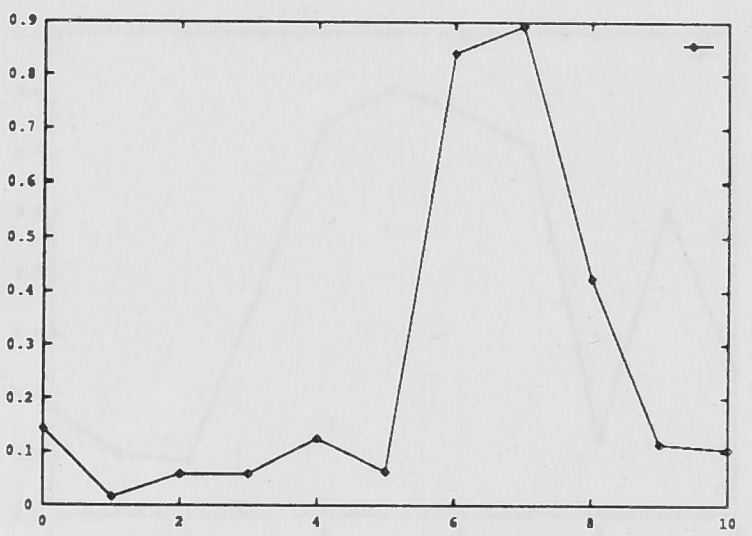
(c)



(d)

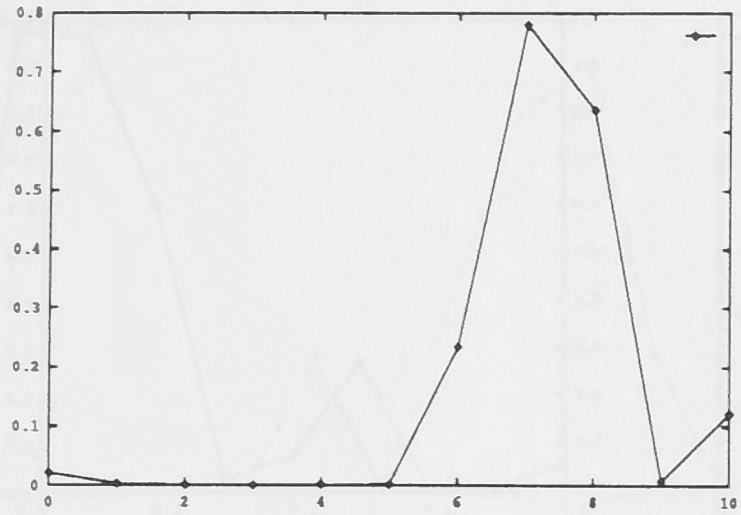


(e)

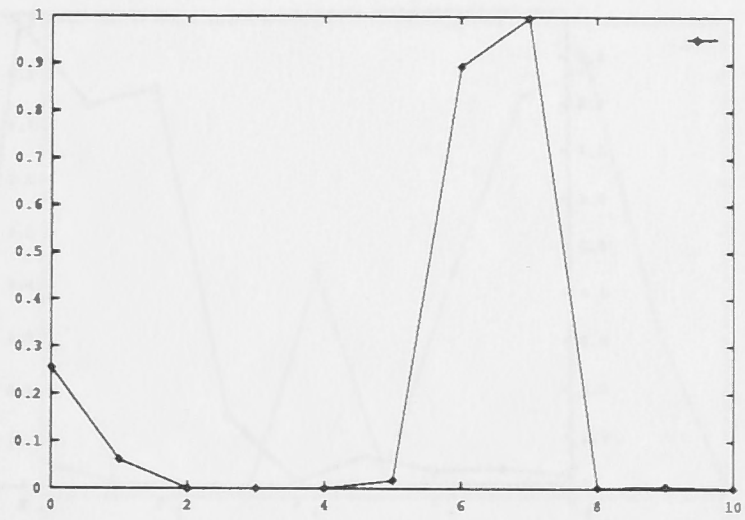


(f)

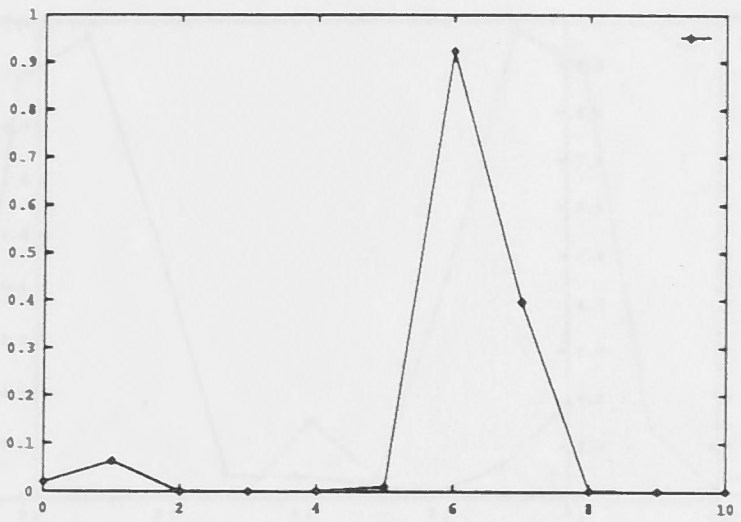
Figure A.36: *Test results of IM's* flat *feature detector tested on IM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



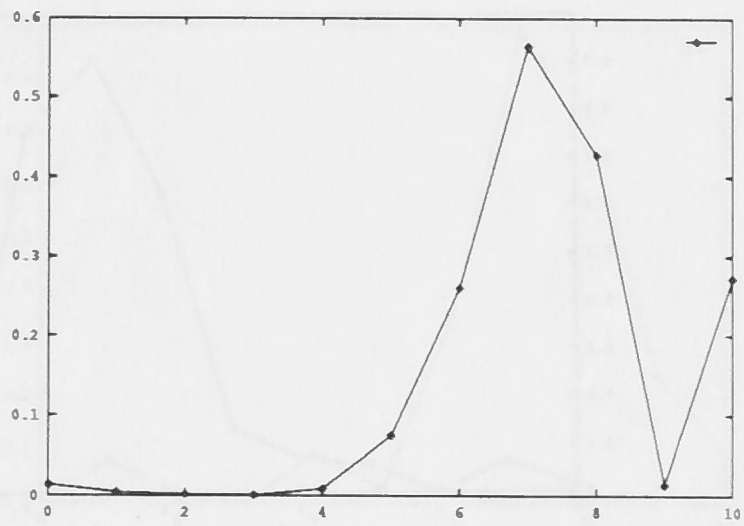
(a)



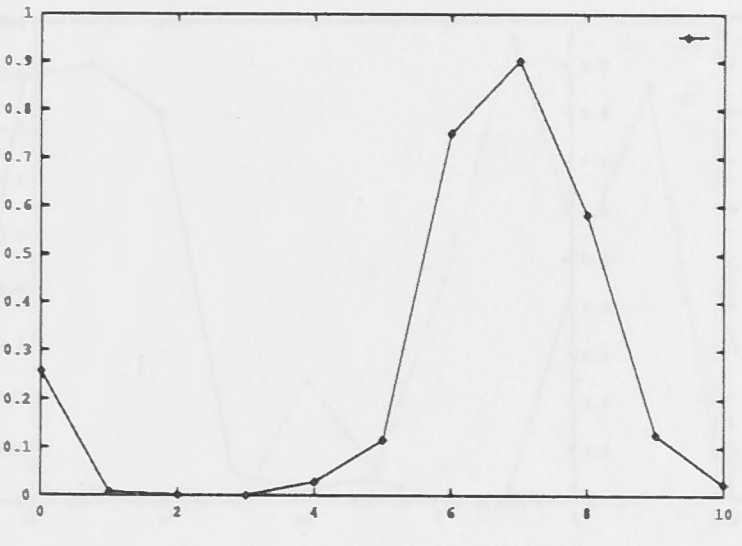
(b)



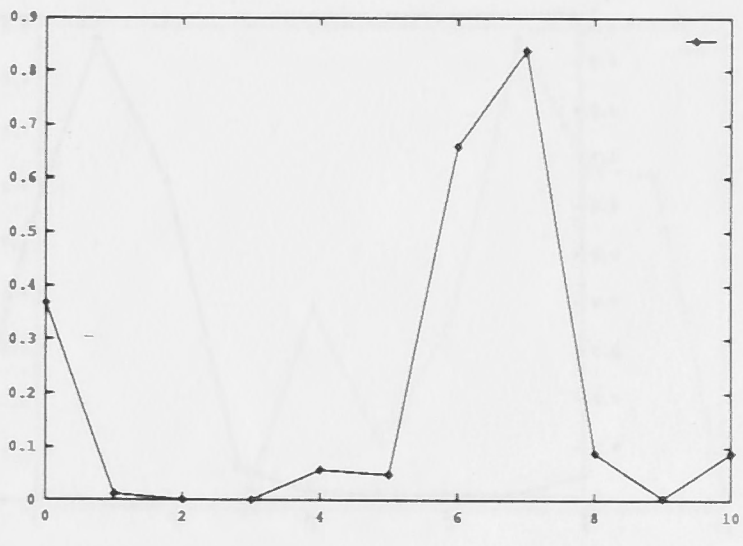
(c)



(d)

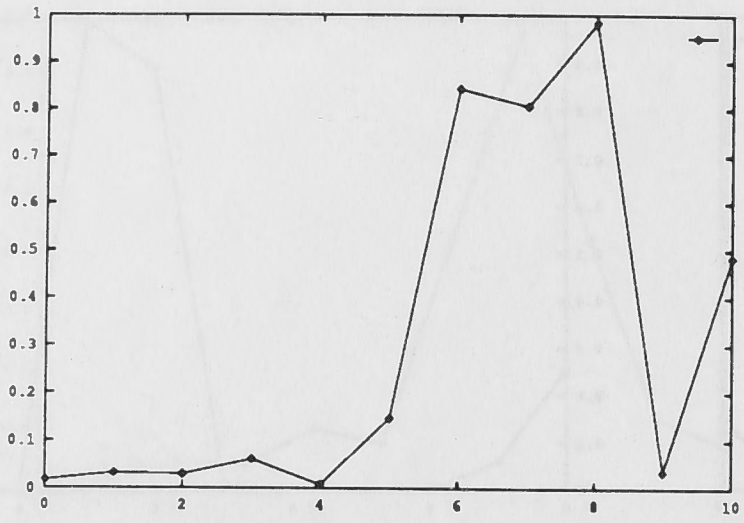


(e)

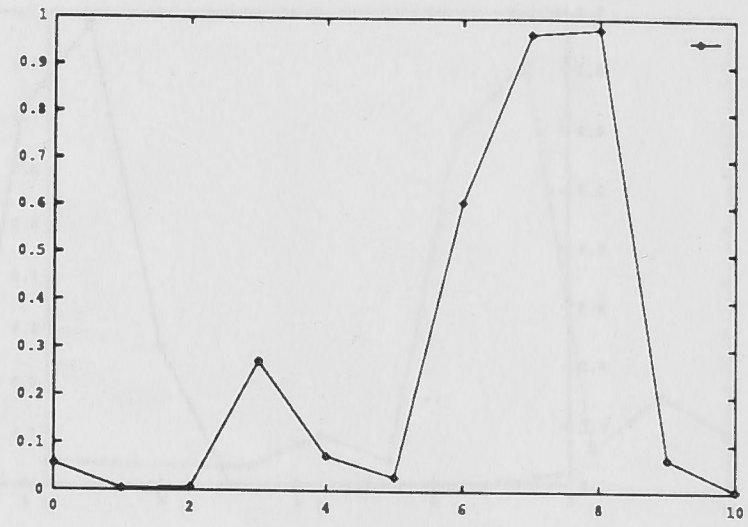


(f)

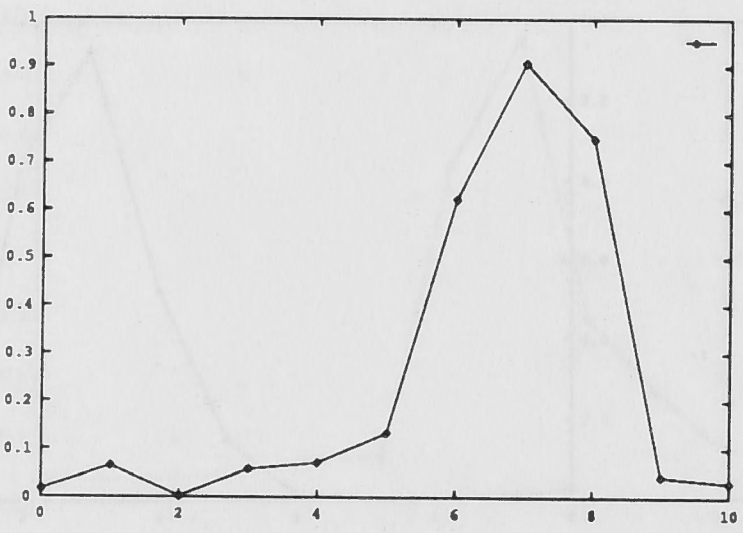
Figure A.37: *Test results of PM's* flat *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



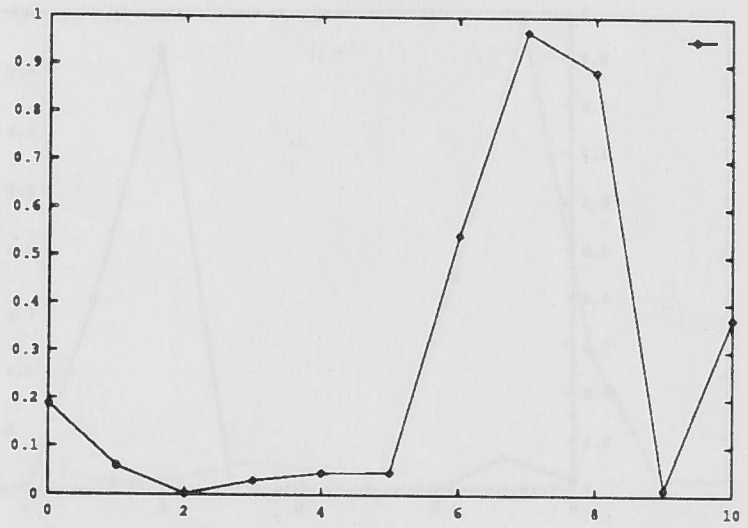
(a)



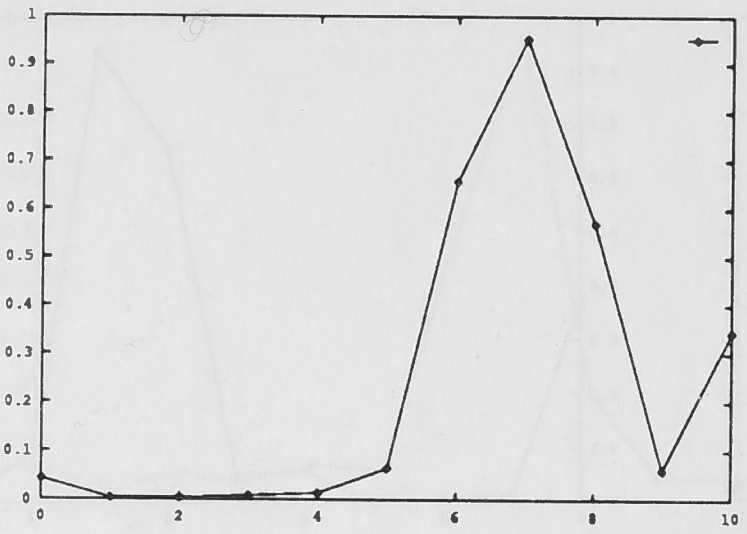
(b)



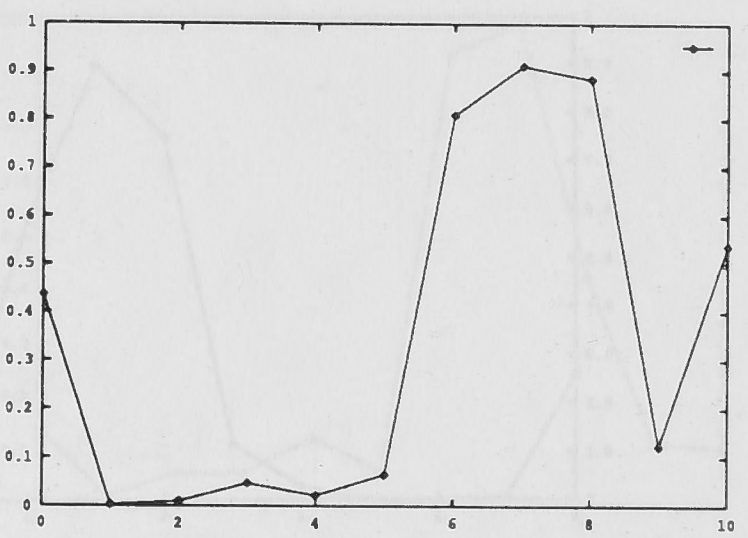
(c)



(d)

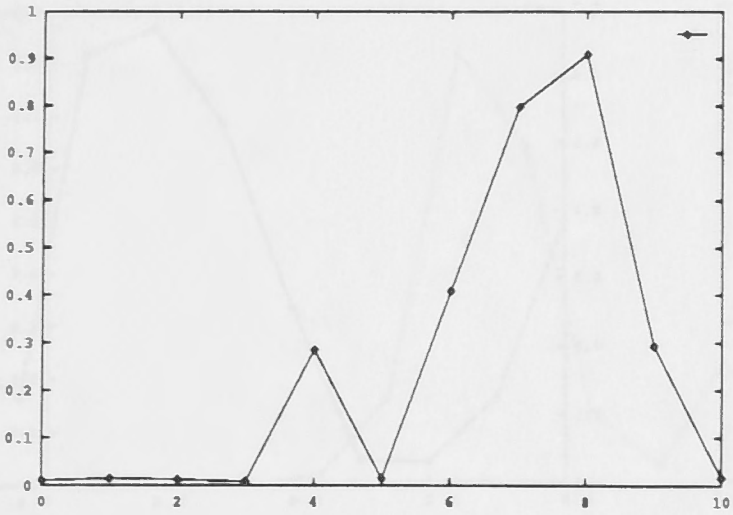


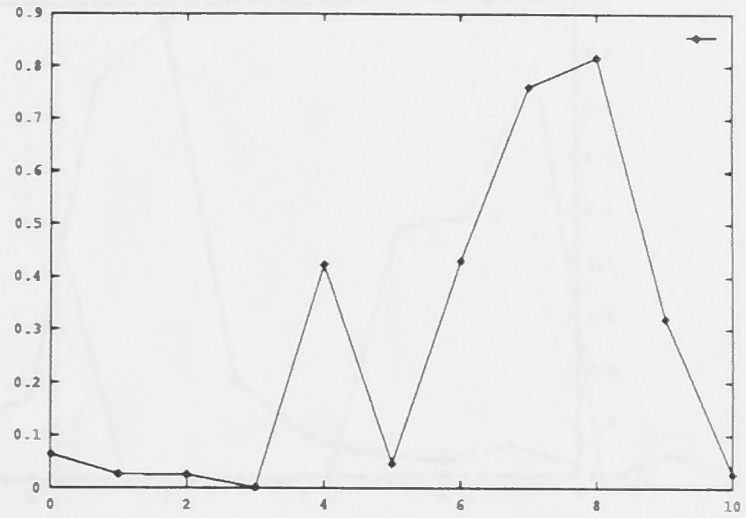
(e)

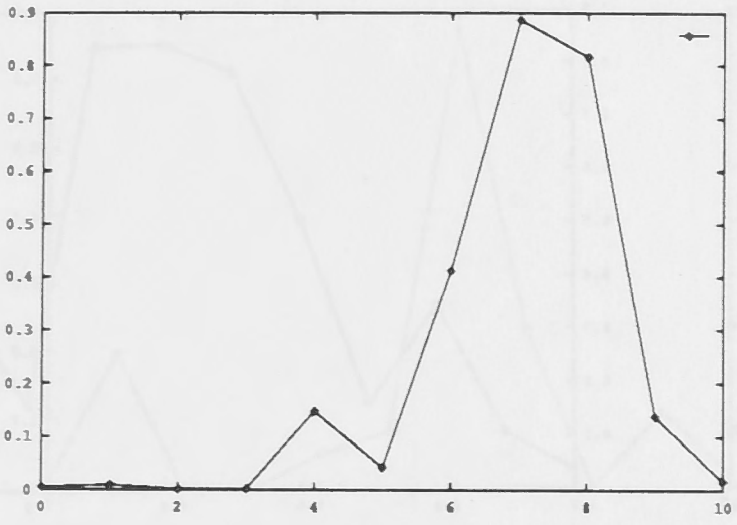


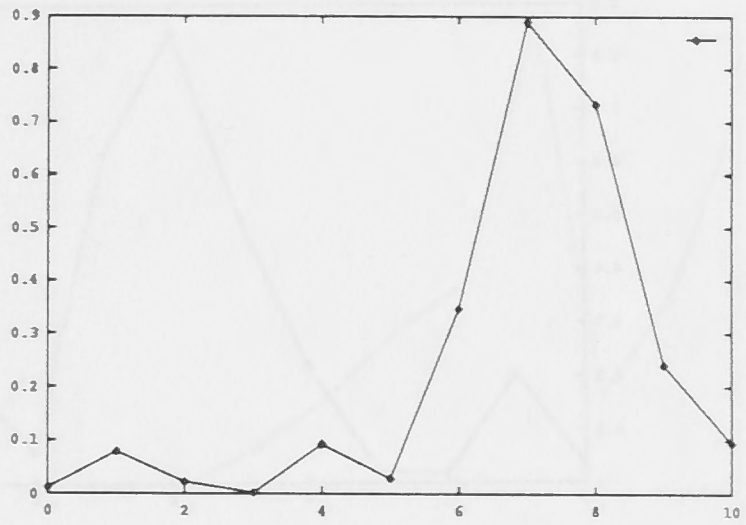
(f)

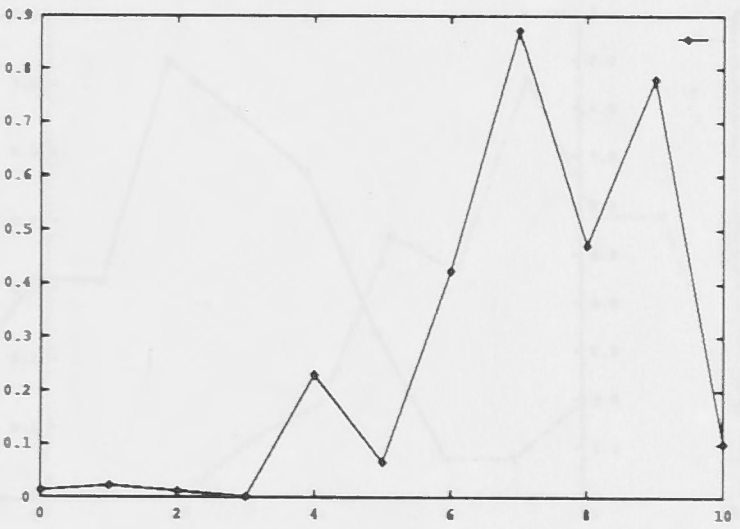
Figure A.38: *Test results of PP's* flat *feature detector tested on PP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].











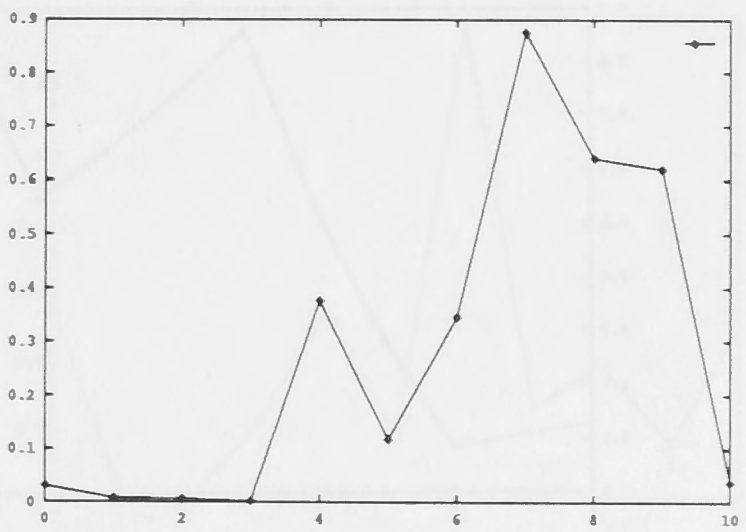
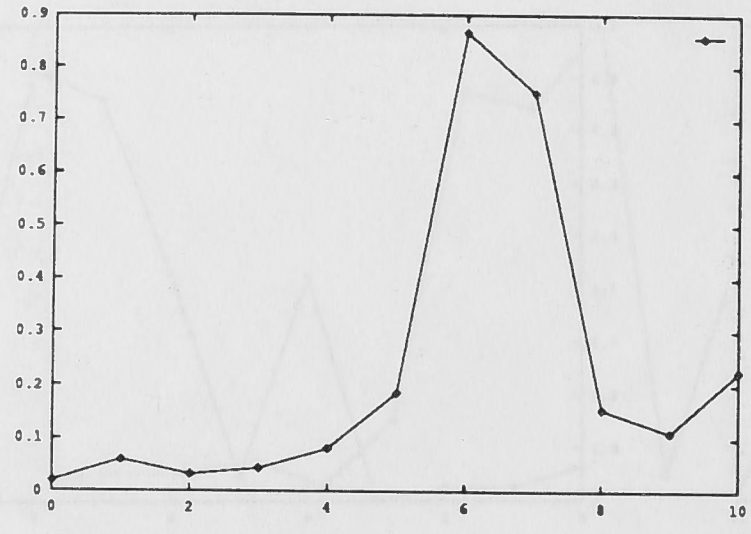
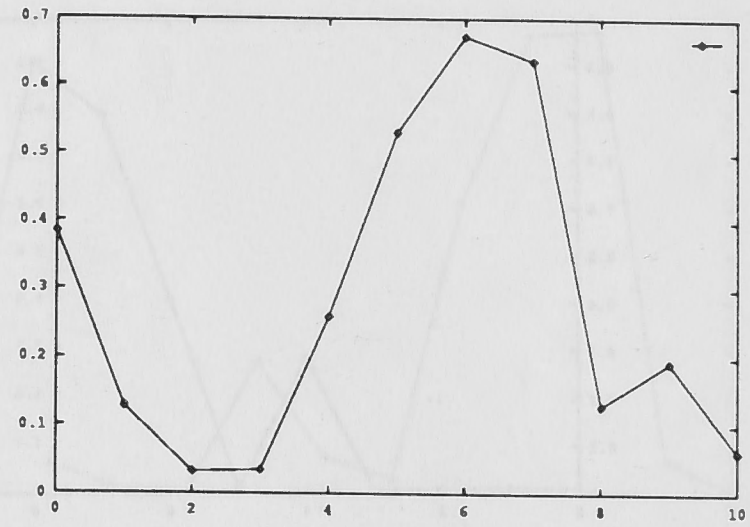


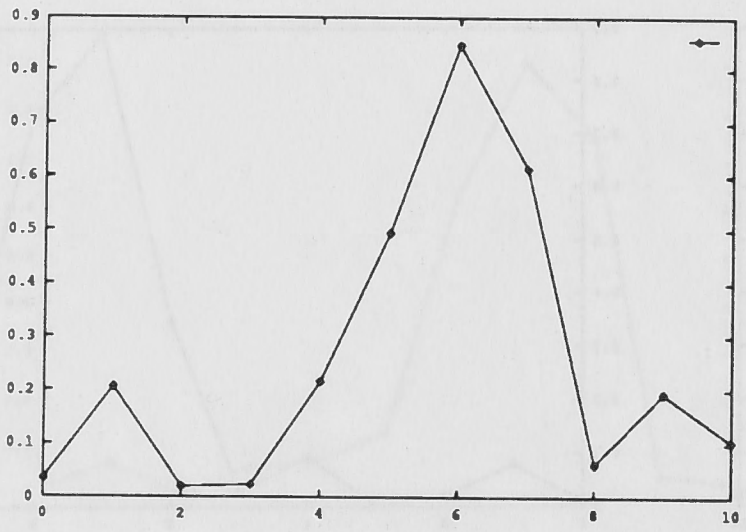
Figure A.39: *Test results of RP's* flat *feature detector tested on RP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



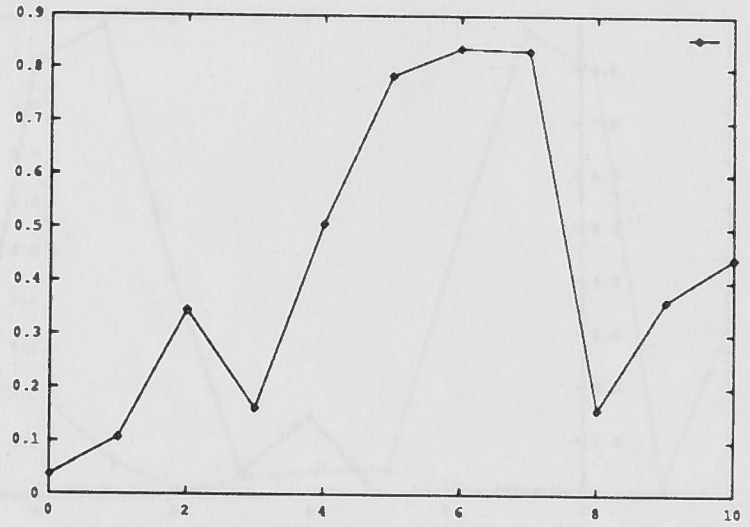
(a)



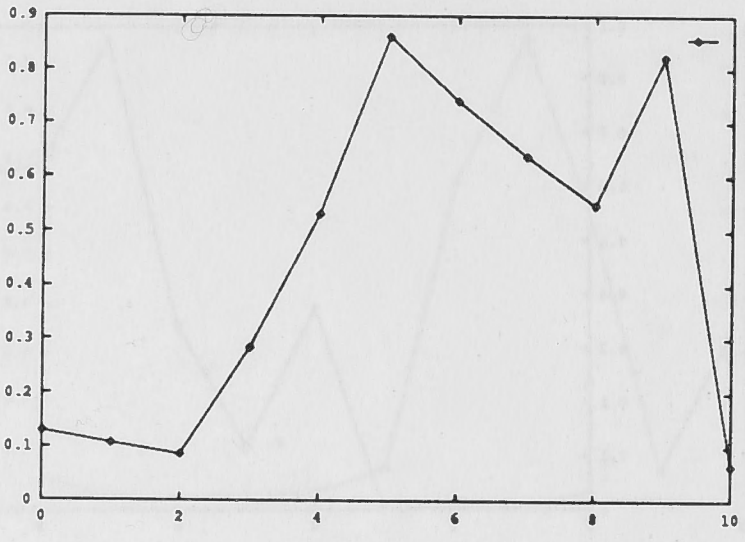
(b)



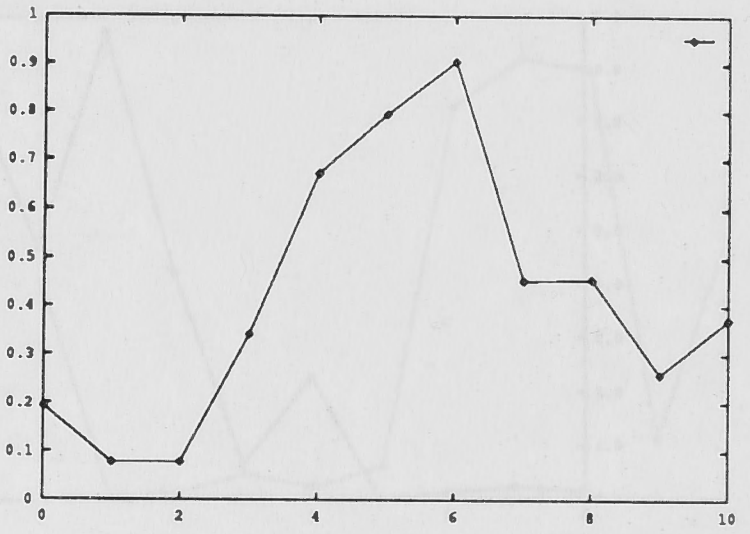
(c)



(d)



(e)



(f)

Figure A.40: *Test results of IM's* flat *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

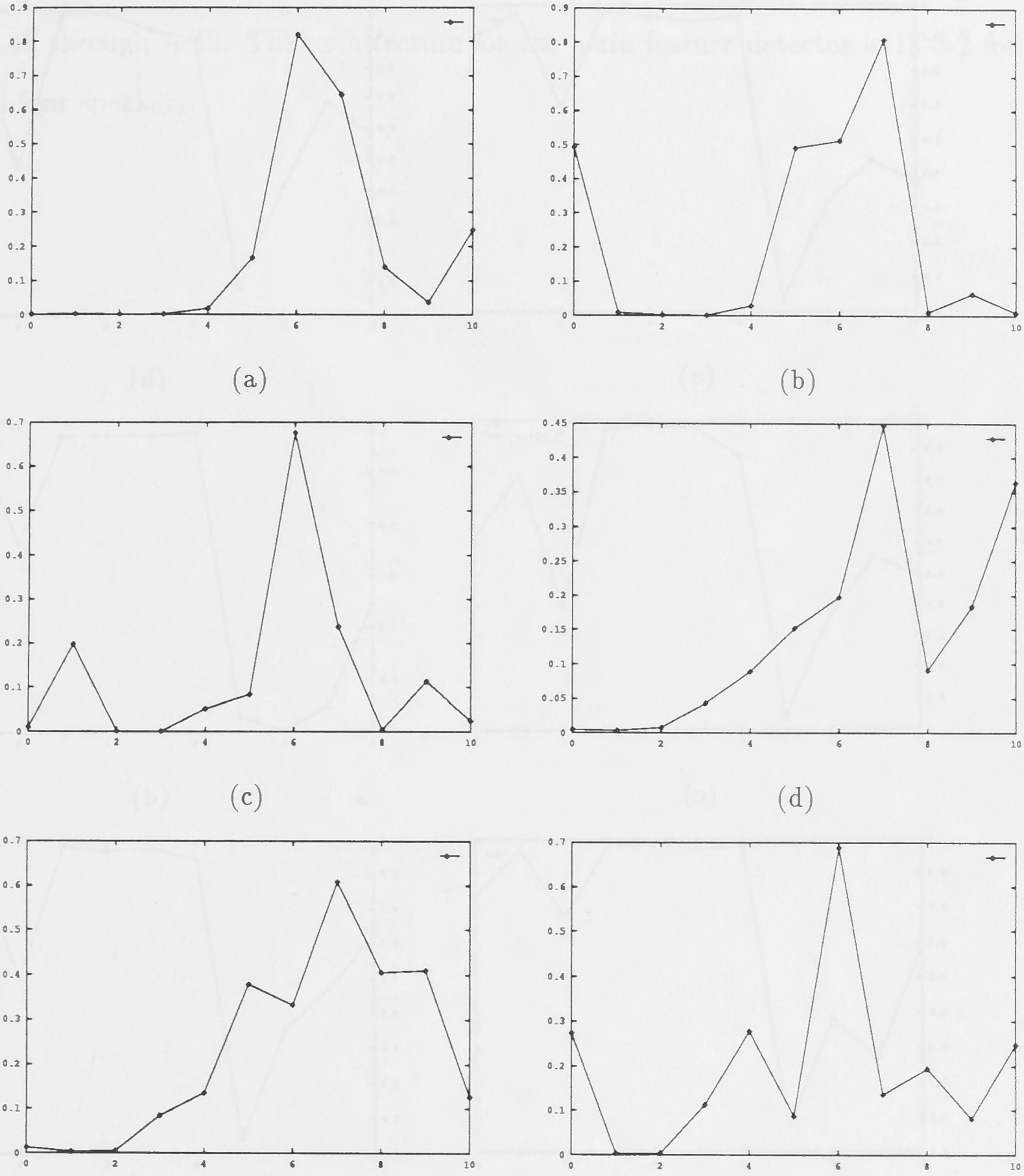
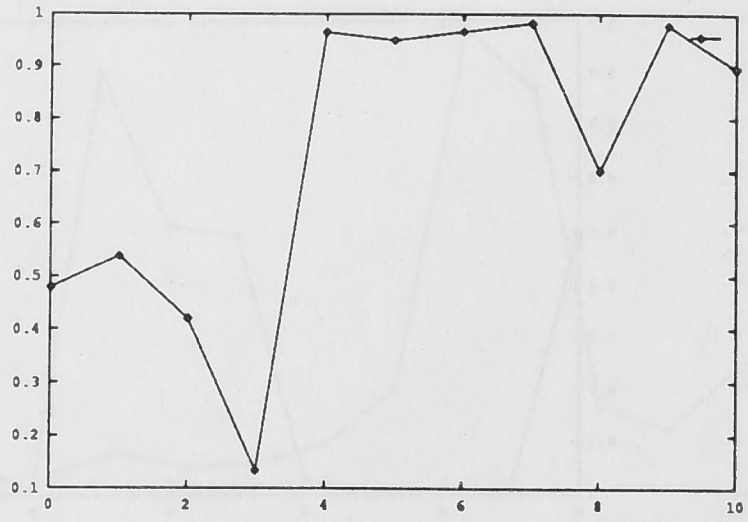
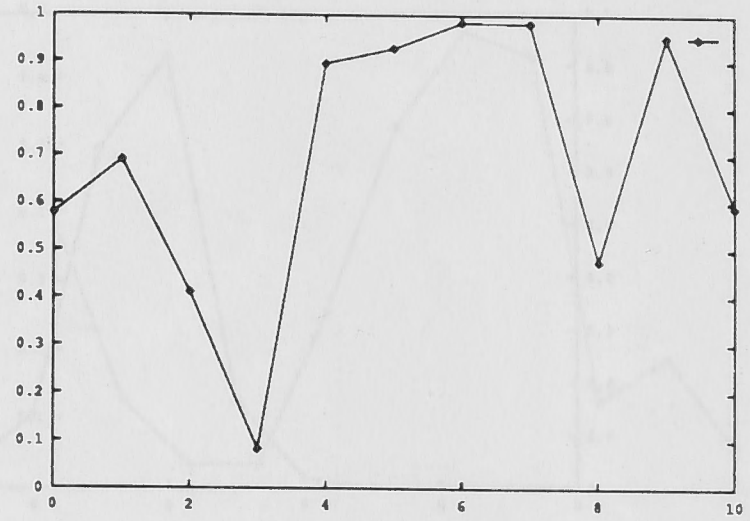


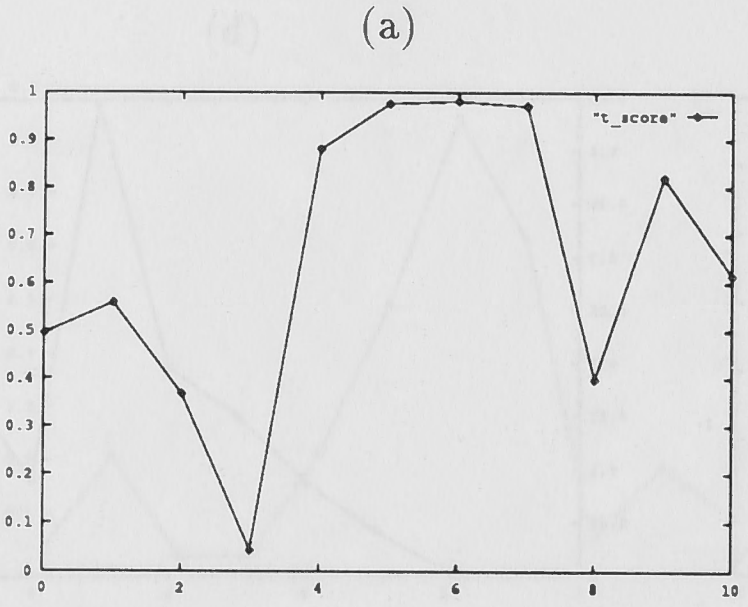
Figure A.41: *Test results of PP's* flat *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ].



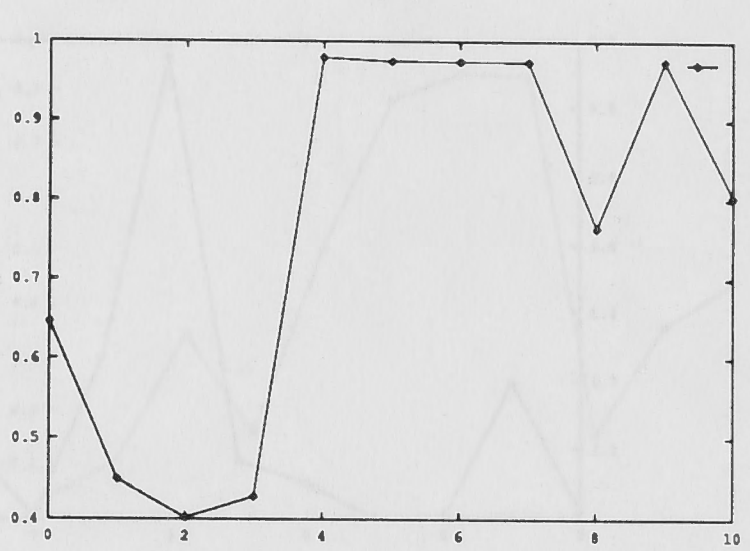
(a)



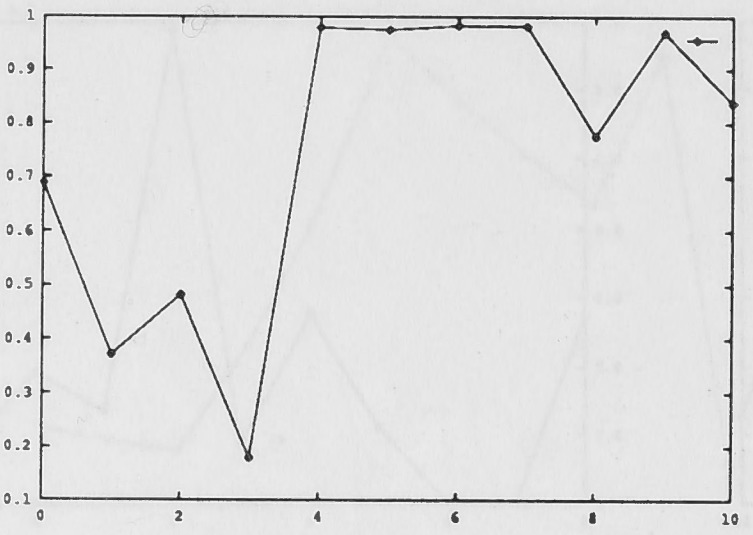
(b)



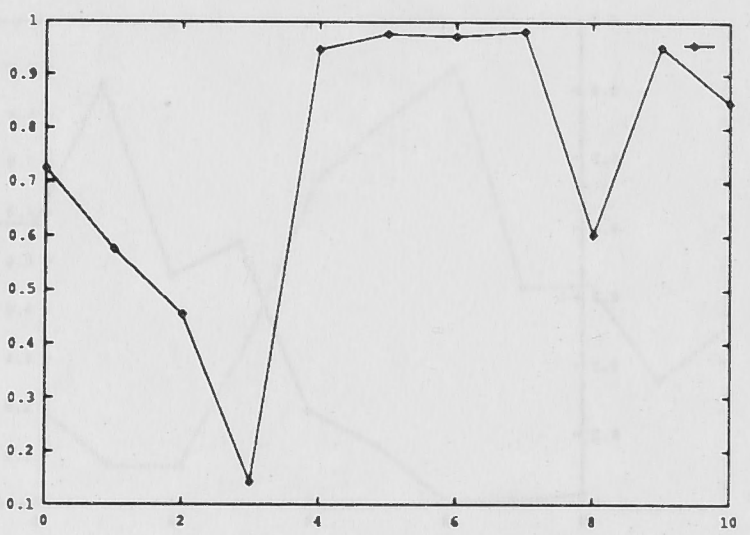
(c)



(d)



(e)

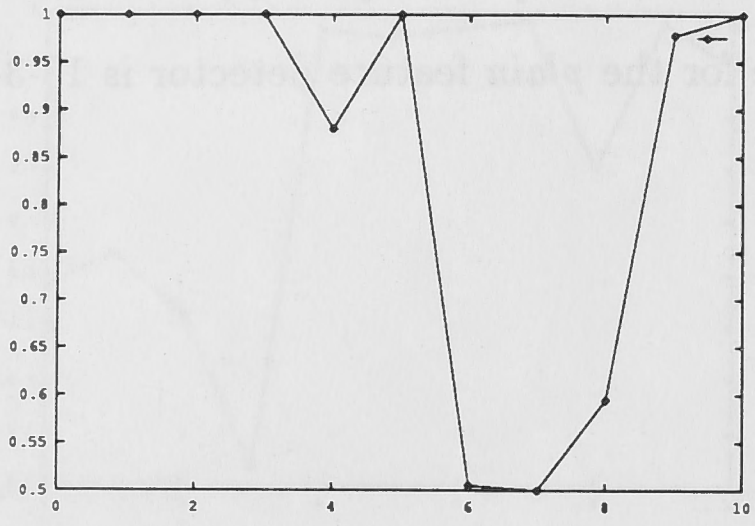


(f)

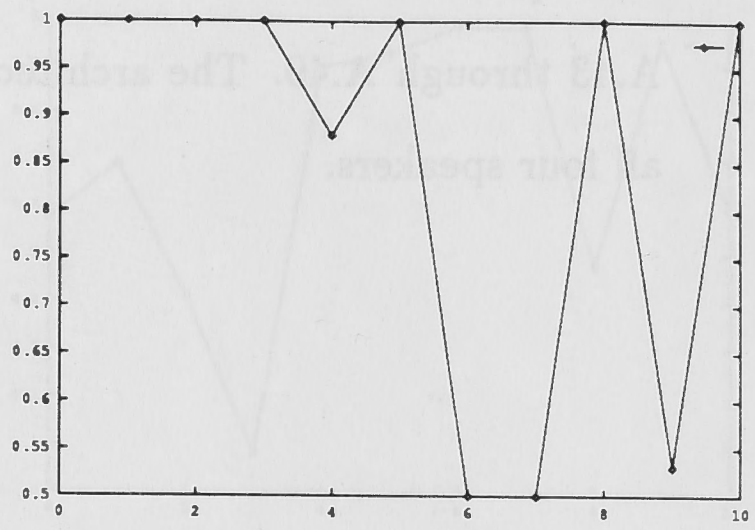
Figure A.42: *Test results of RP's* flat *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

A.6 Feature "Plain"

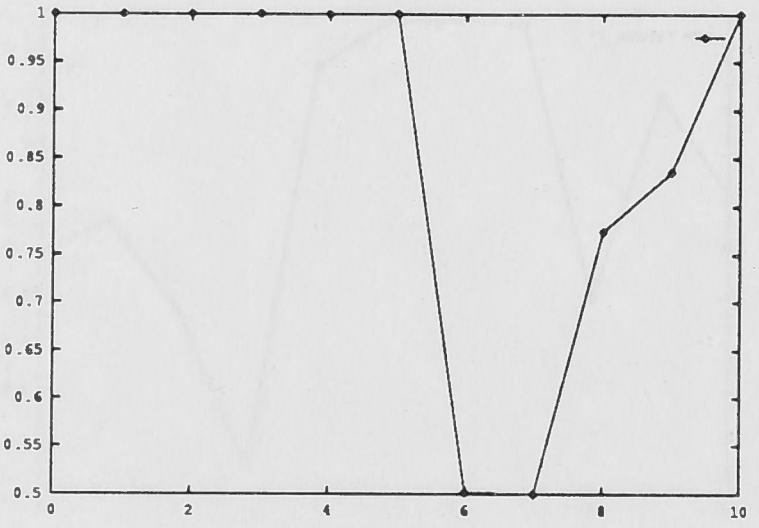
This section presents a complete set of test results of feature *plain* detector for each speaker tested on themselves and on PM respectively, as reported in Figures A.43 through A.49. The architecture for the *plain* feature detector is 13-3-2 for all four speakers.



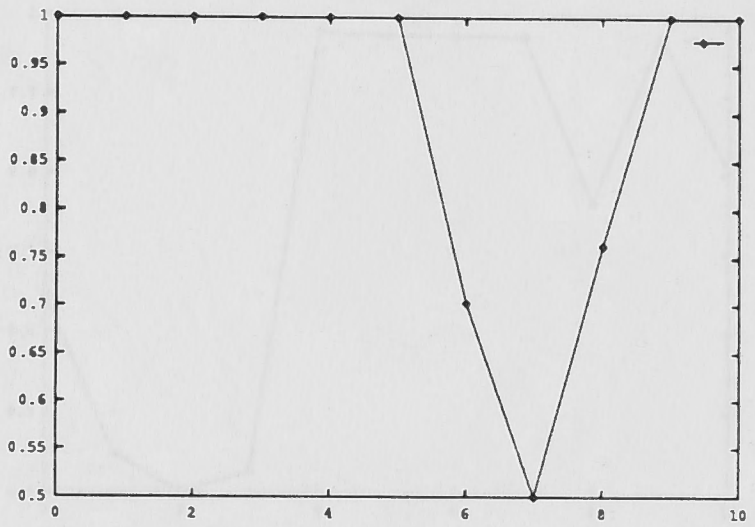
(a)



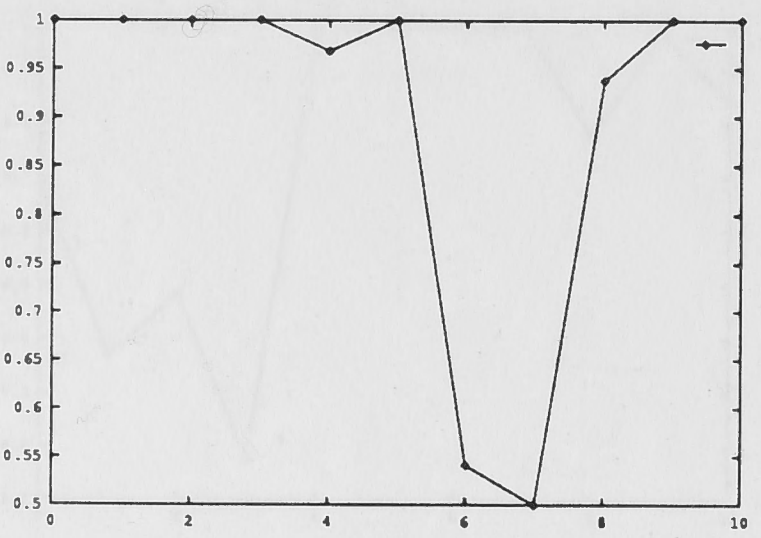
(b)



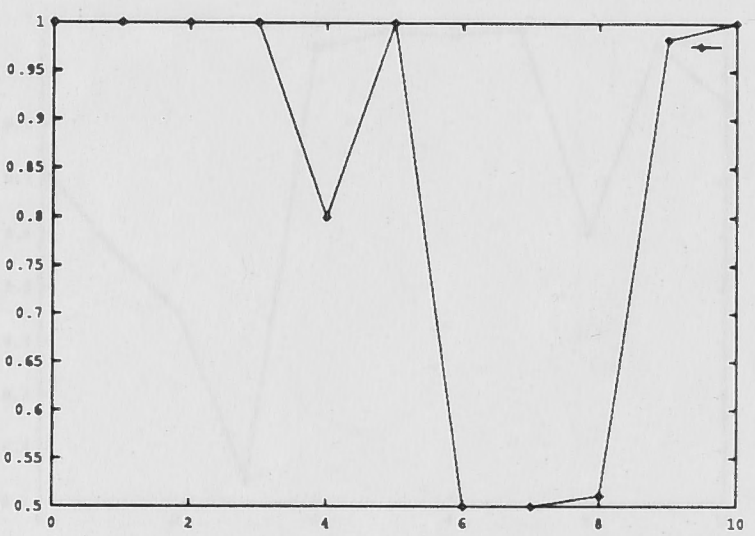
(c)



(d)



(e)



(f)

Figure A.43: *Test results of IM's* plain *feature detector tested on IM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

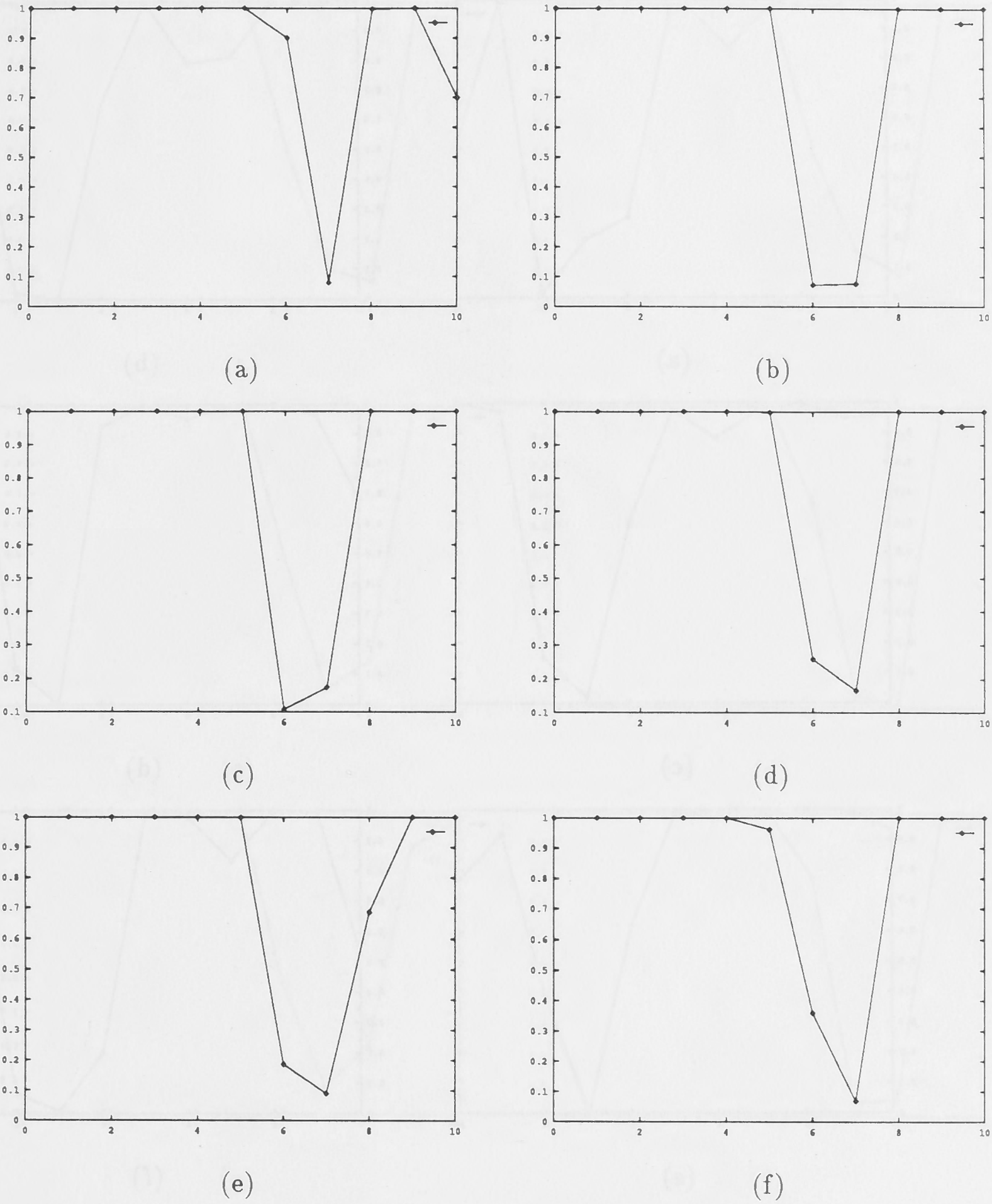
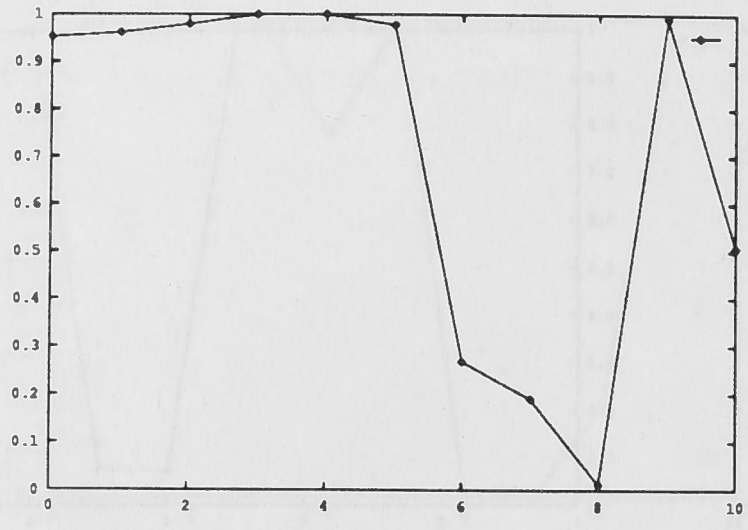
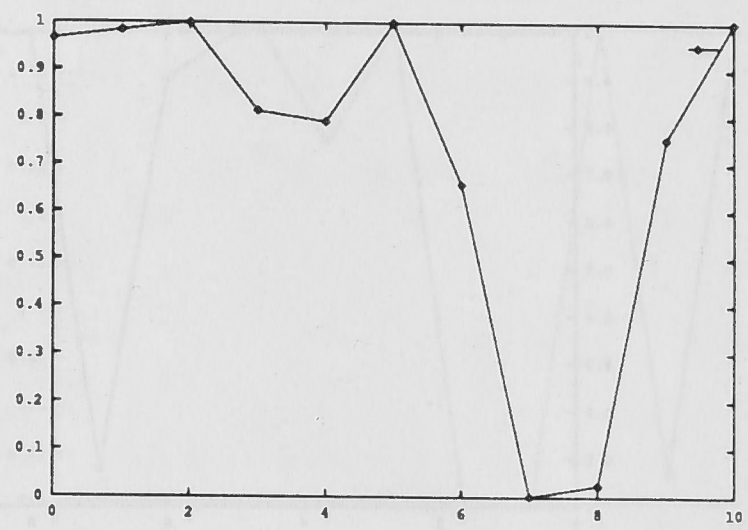


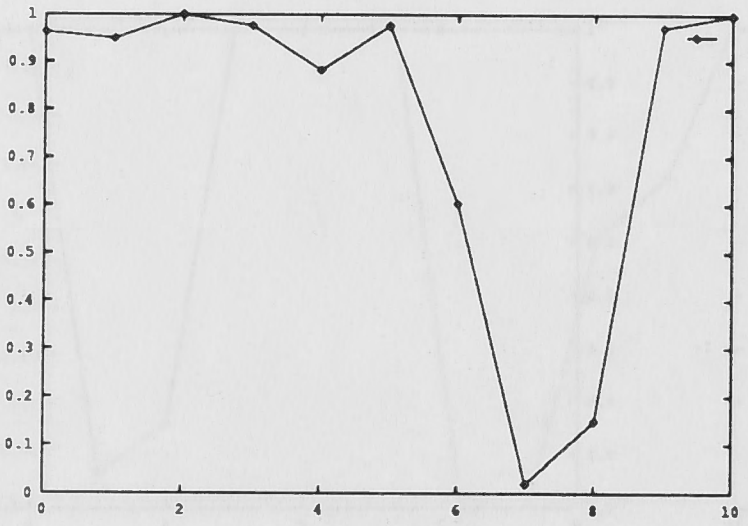
Figure A.44: *Test results of PM's* plain *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



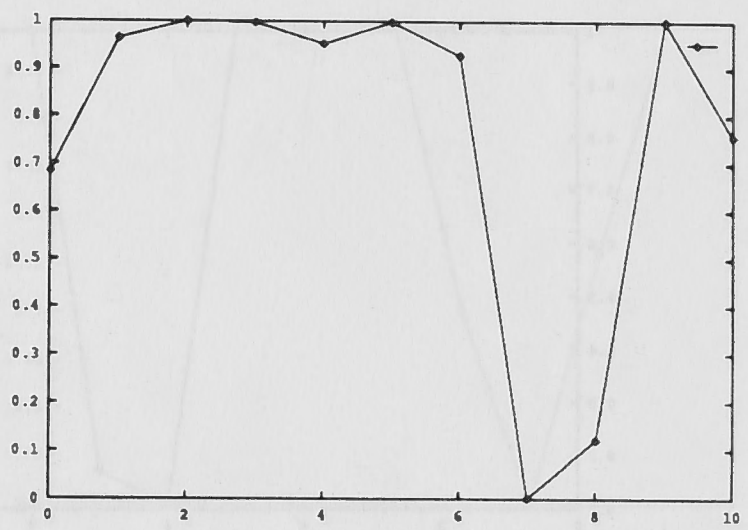
(a)



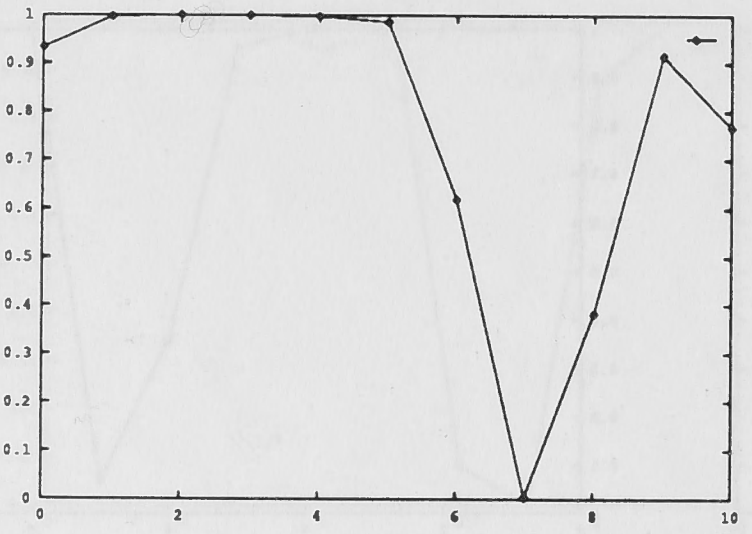
(b)



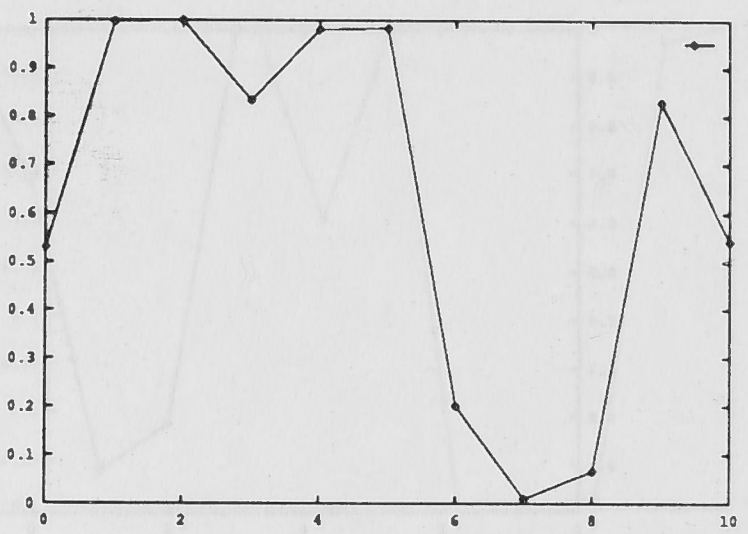
(c)



(d)

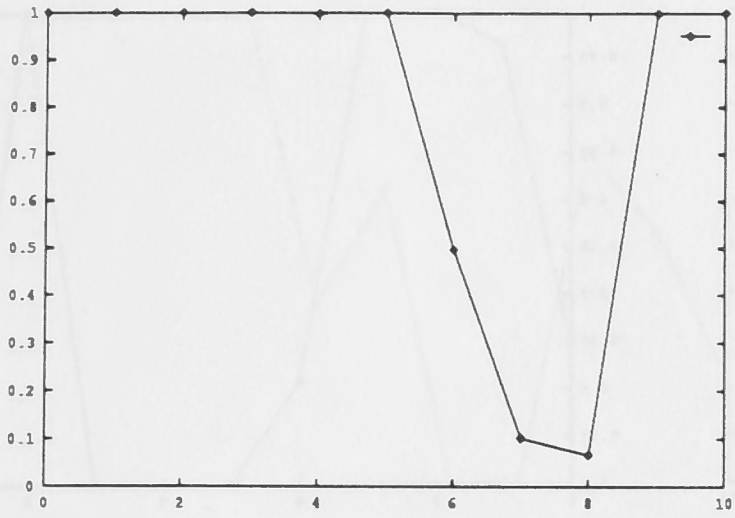


(e)

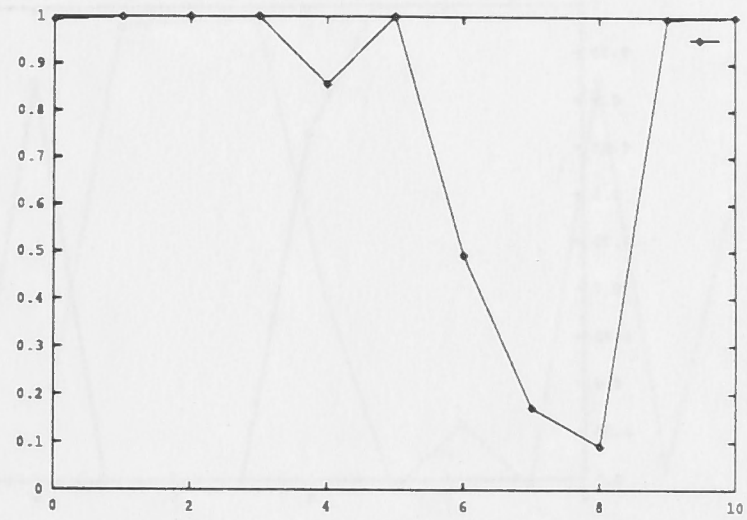


(f)

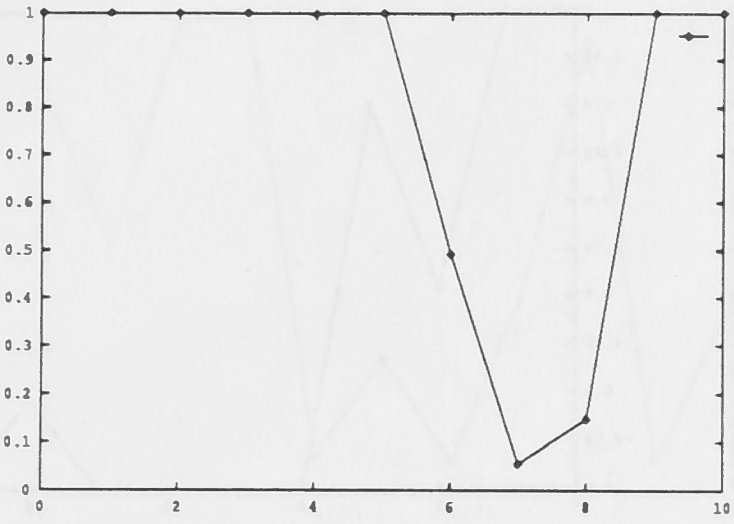
Figure A.45: *Test results of PP's* plain *feature detector tested on PP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



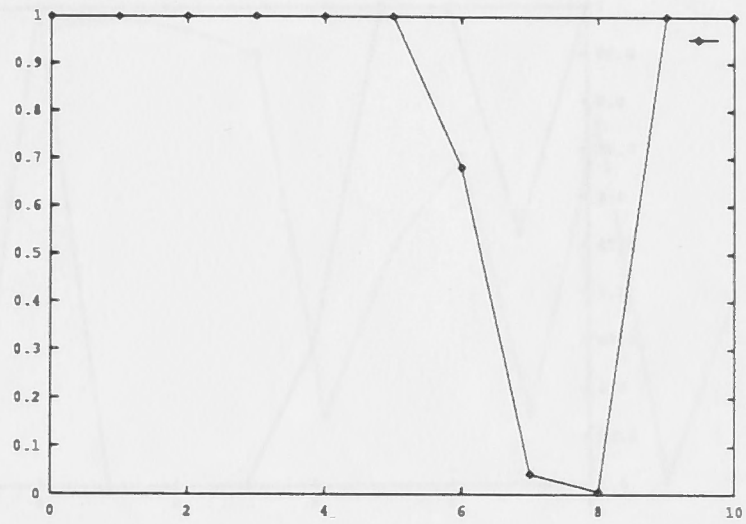
(a)



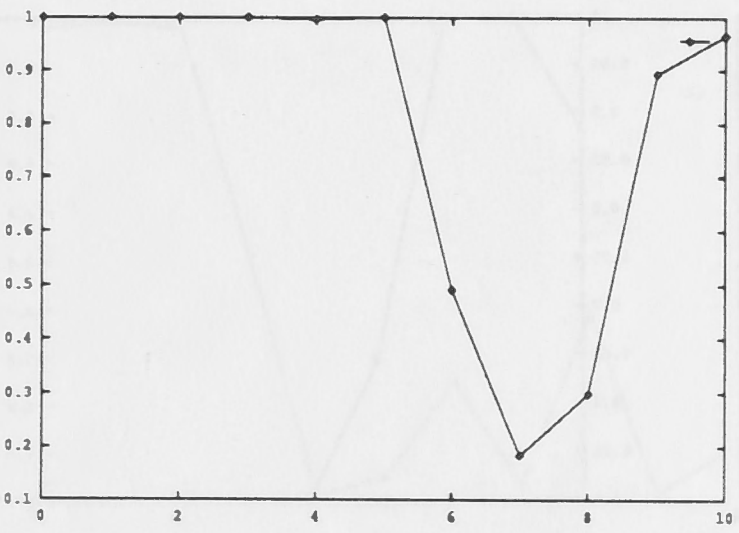
(b)



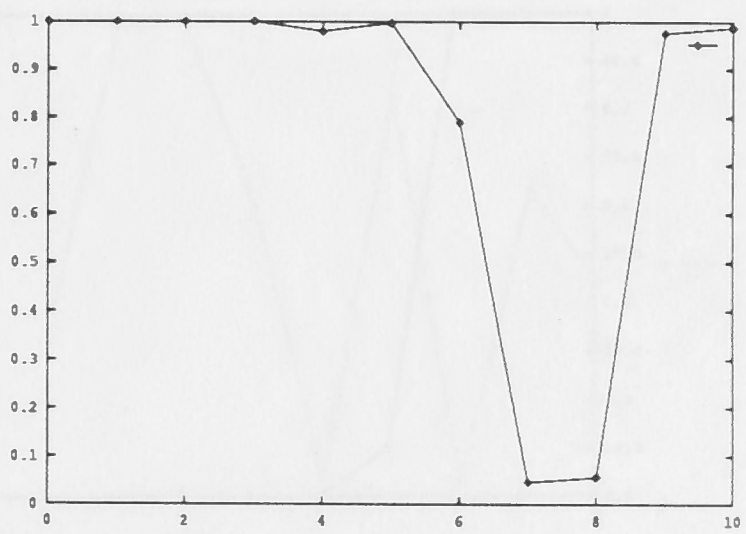
(c)



(d)

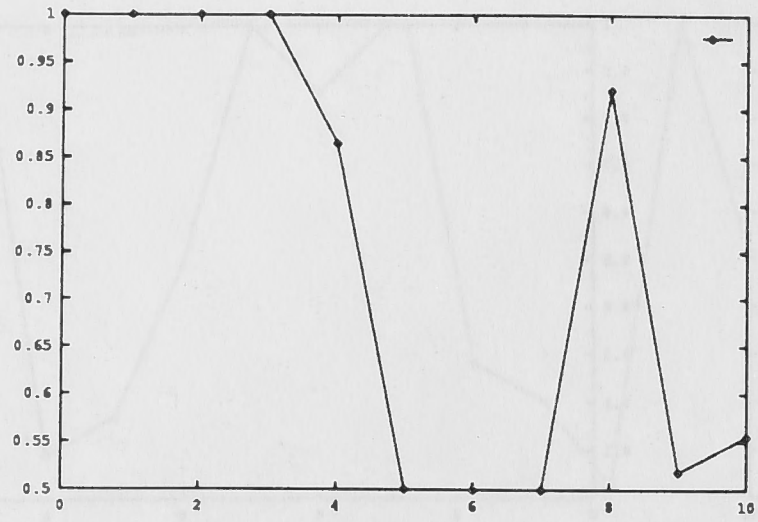


(e)

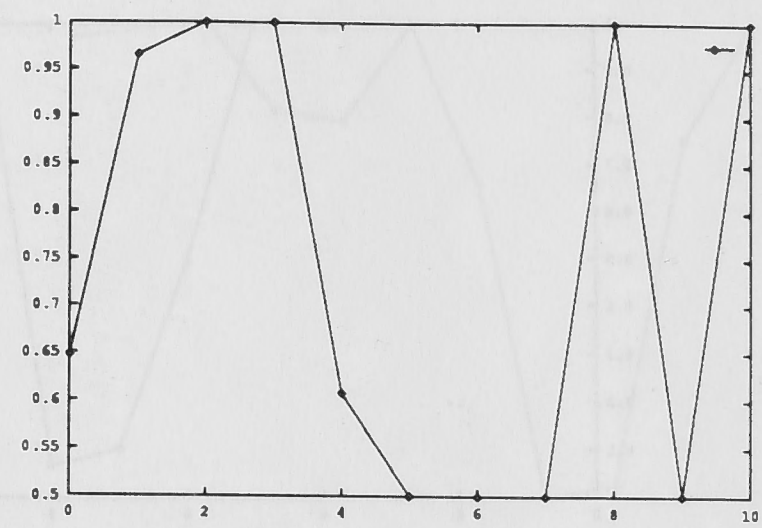


(f)

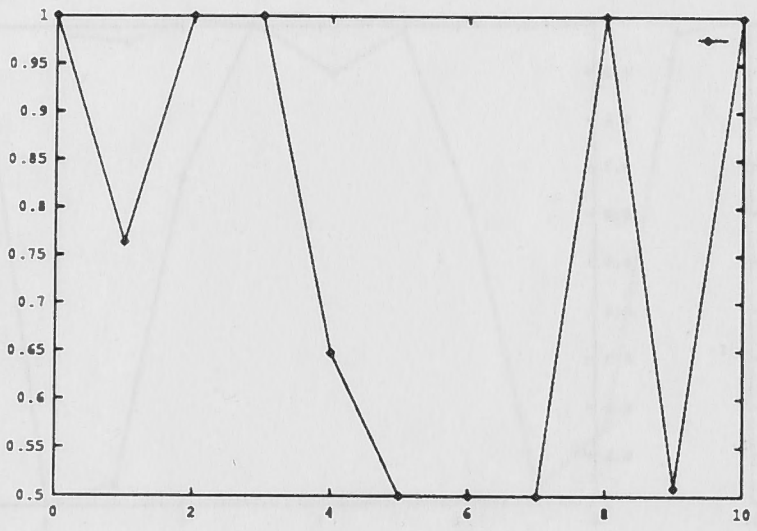
Figure A.46: *Test results of RP's* plain *feature detector tested on RP's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



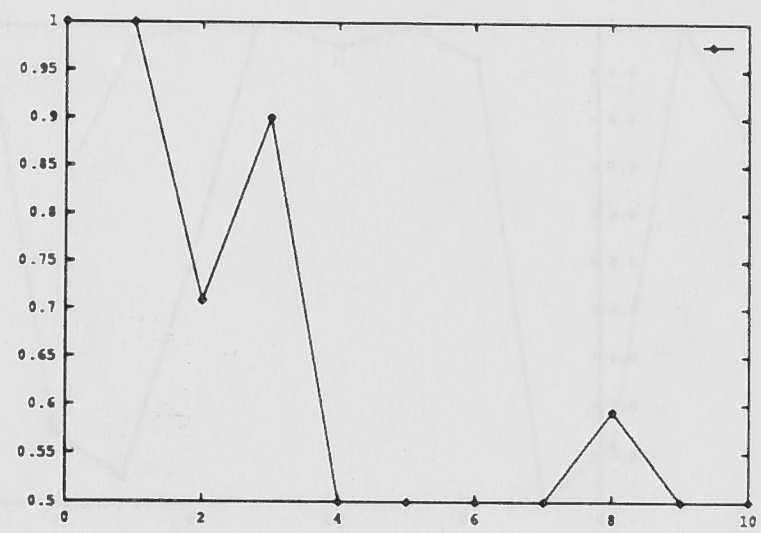
(a)



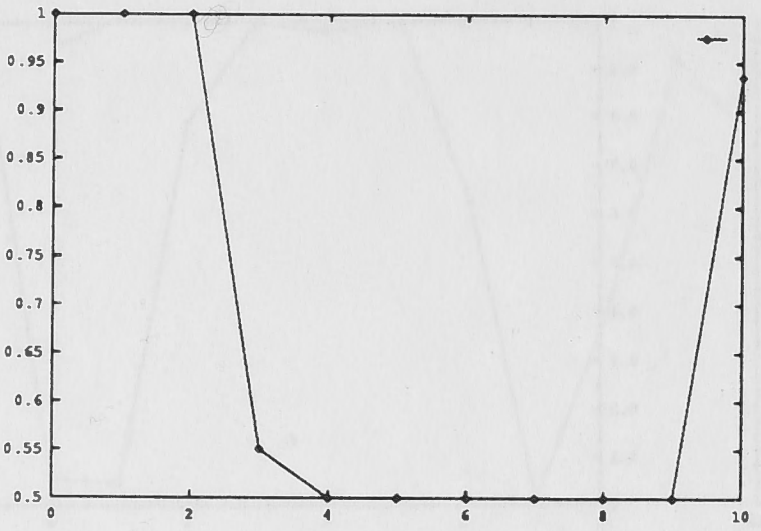
(b)



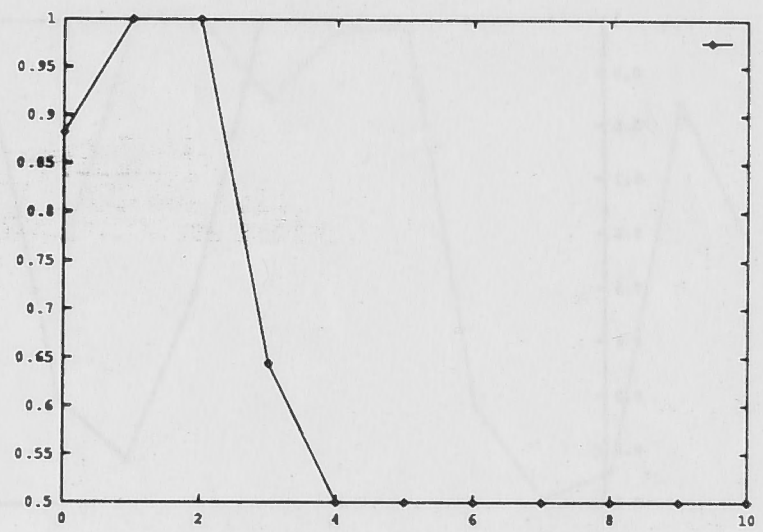
(c)



(d)

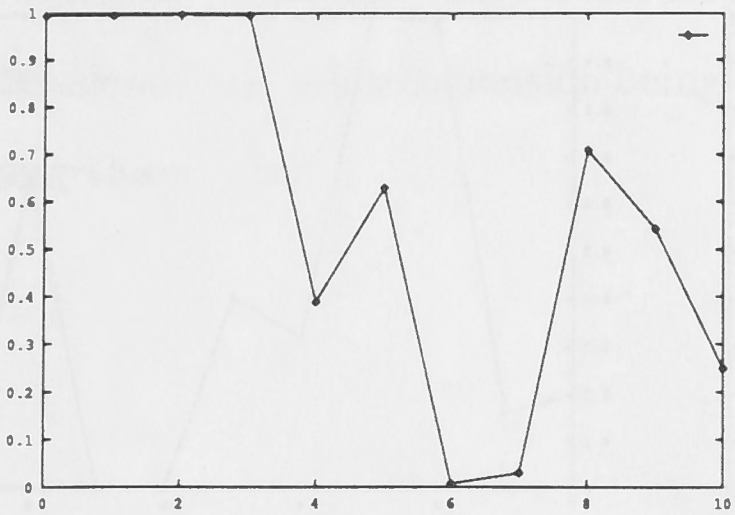


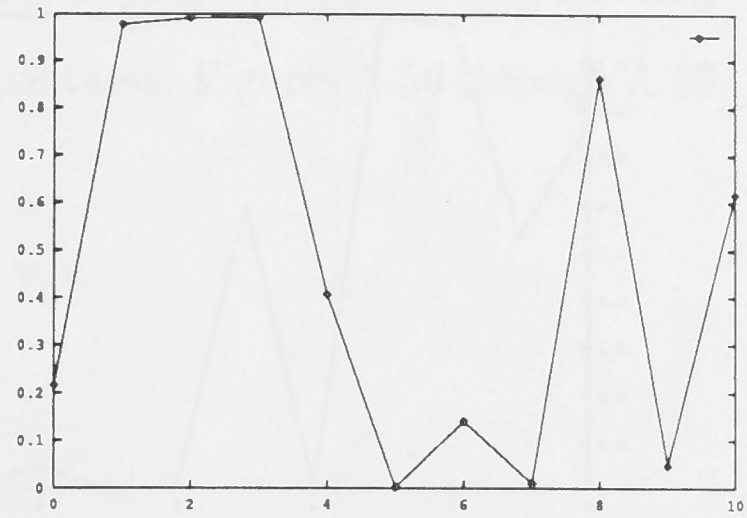
(e)

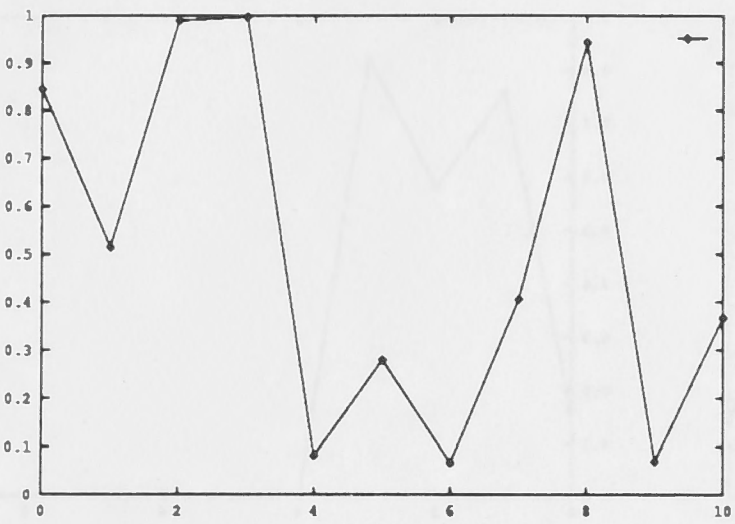


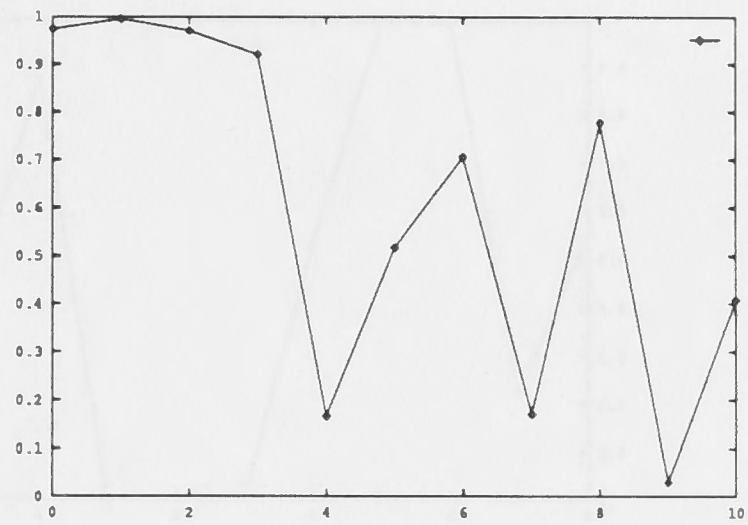
(f)

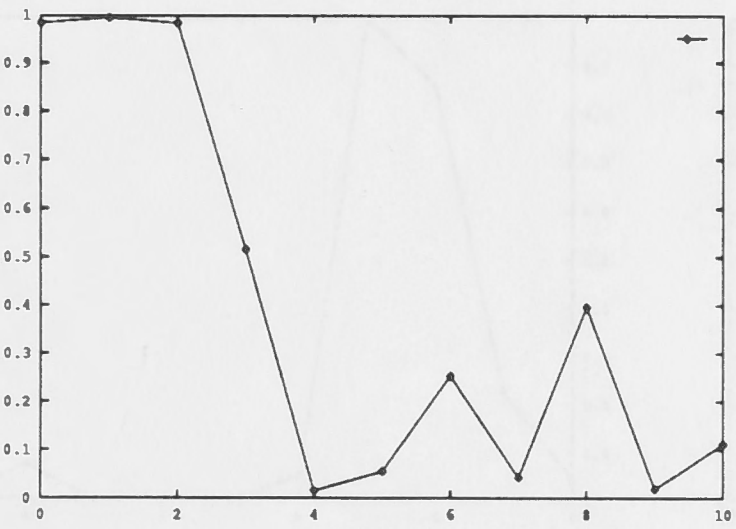
Figure A.47: *Test results of IM's* plain *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].











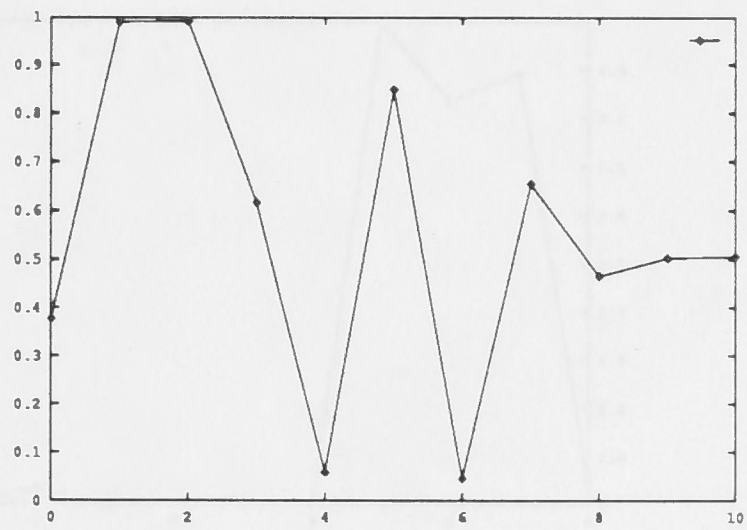
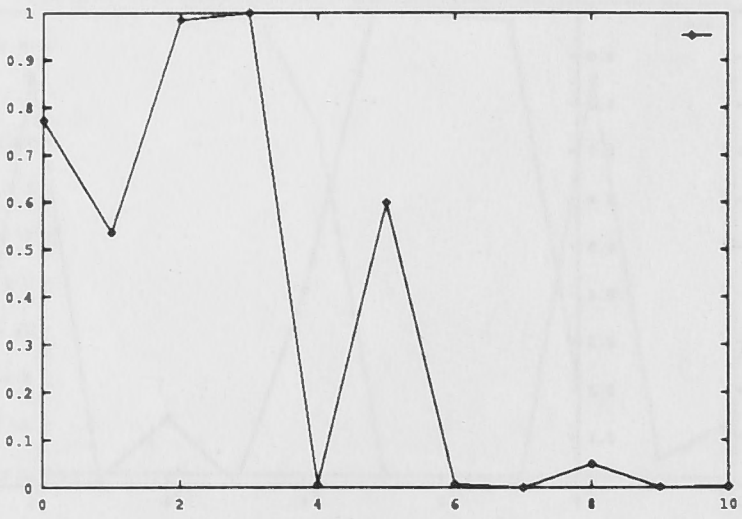
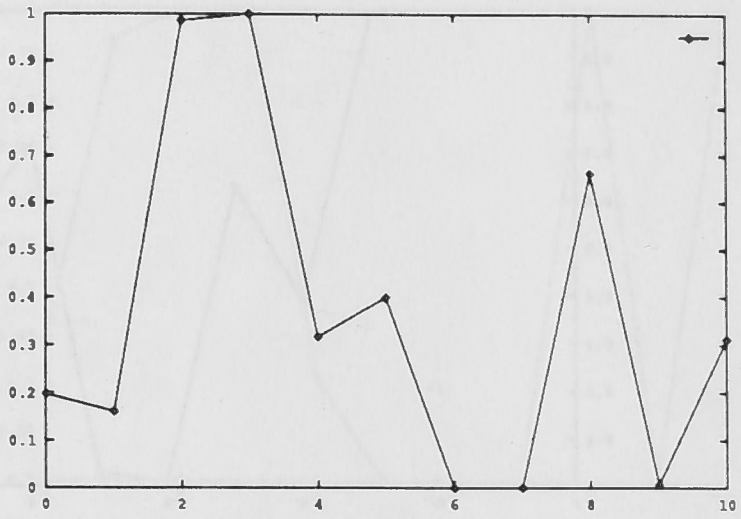


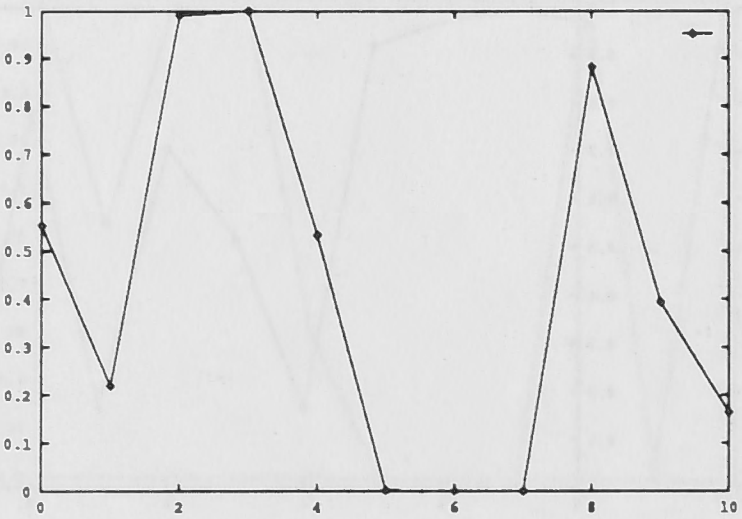
Figure A.48: *Test results of PP's* plain *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].



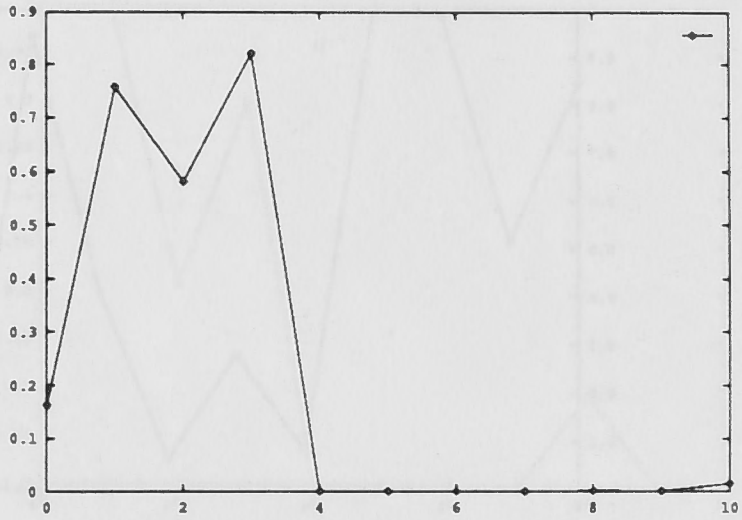
(a)



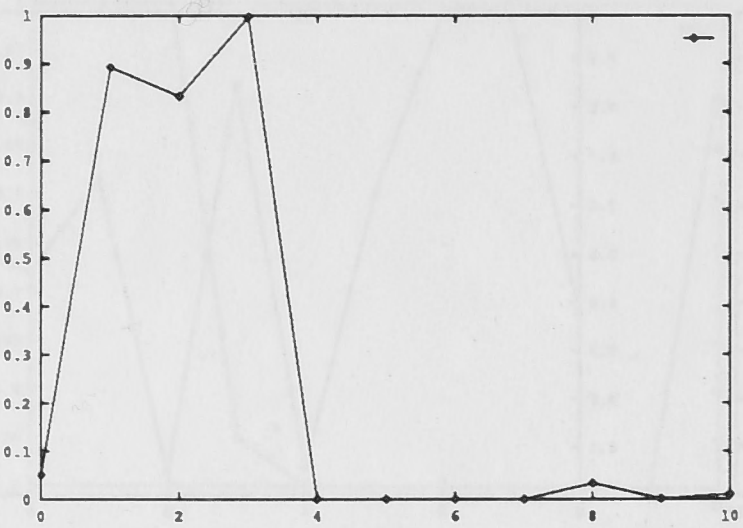
(b)



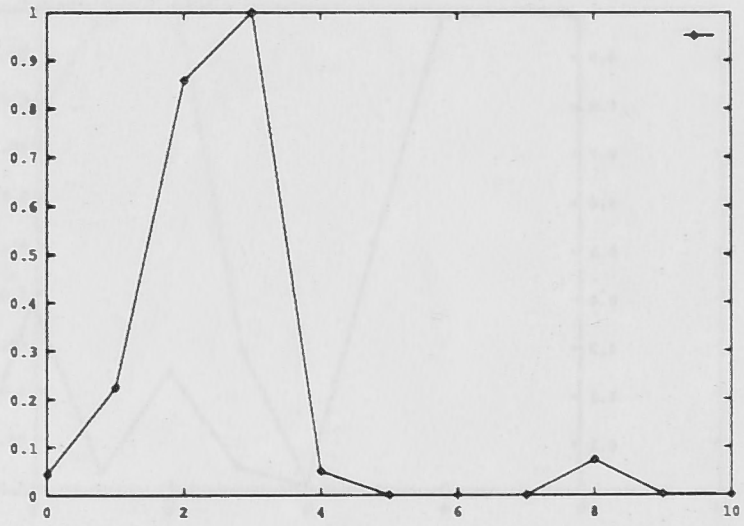
(c)



(d)



(e)

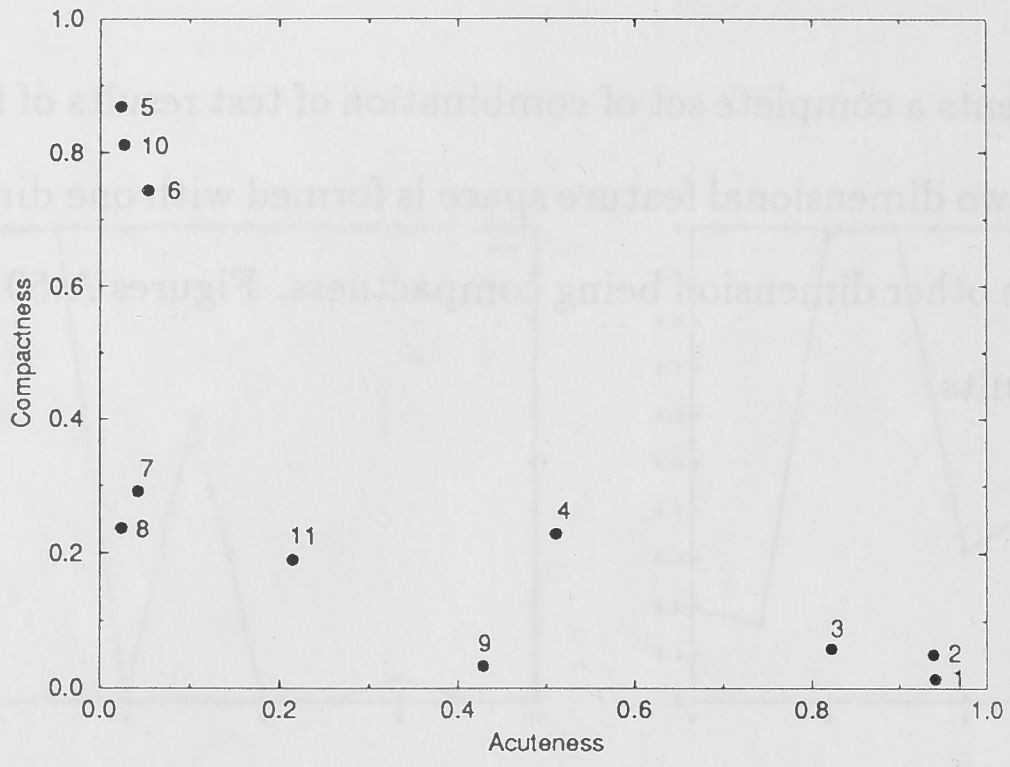


(f)

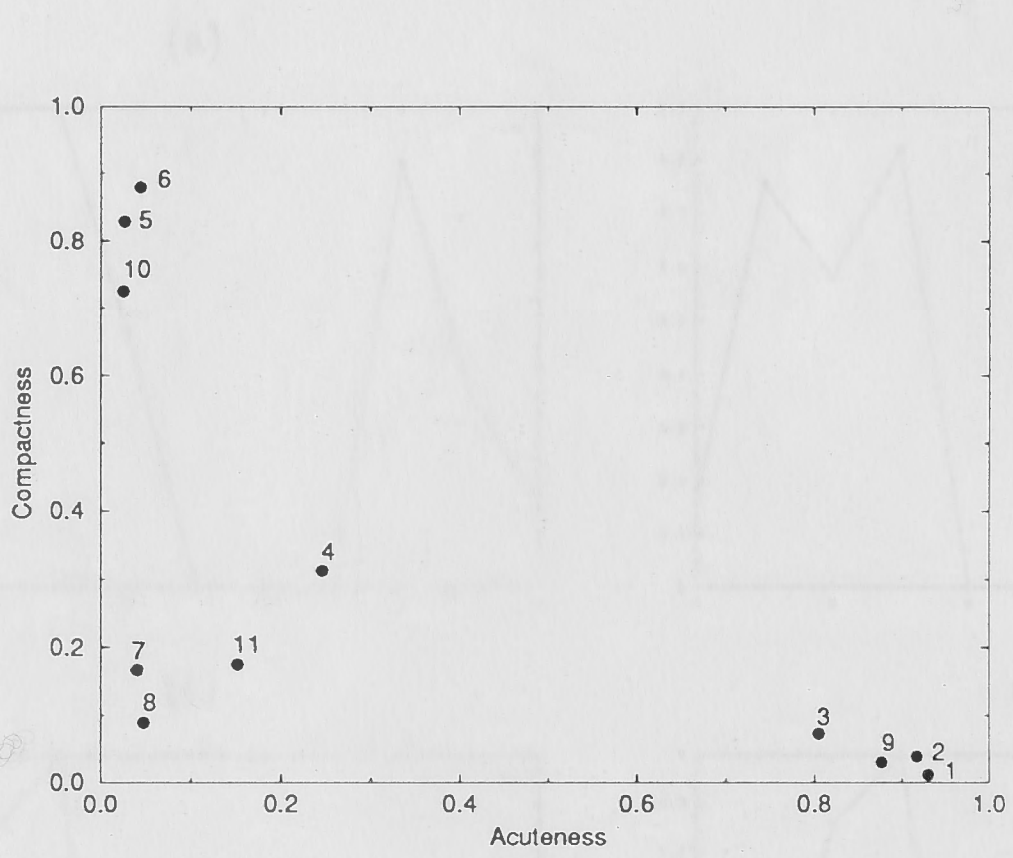
Figure A.49: *Test results of RP's* plain *feature detector tested on PM's 11 pseudo steady-state vowels in the context of 6 stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**].

A.7 Acuteness versus Compactness

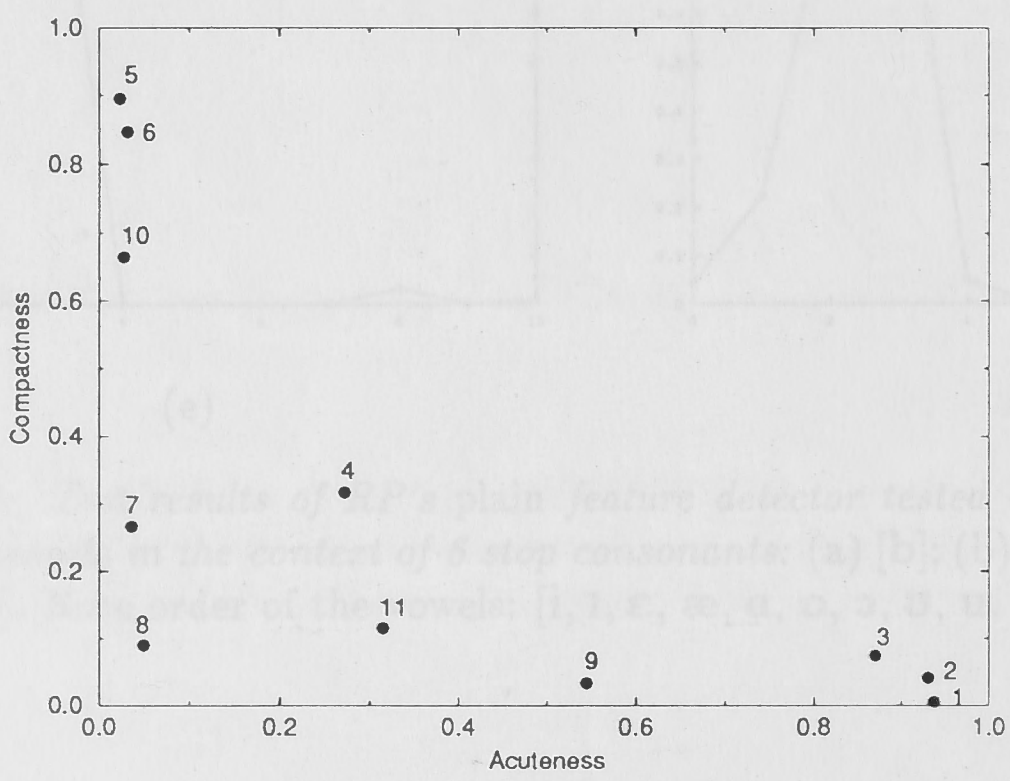
This section presents a complete set of combination of test results of features *acute* and *compact*. A two dimensional feature space is formed with one dimension being acuteness and the other dimension being compactness. Figures A.50 through A.53 present these results.



(a)

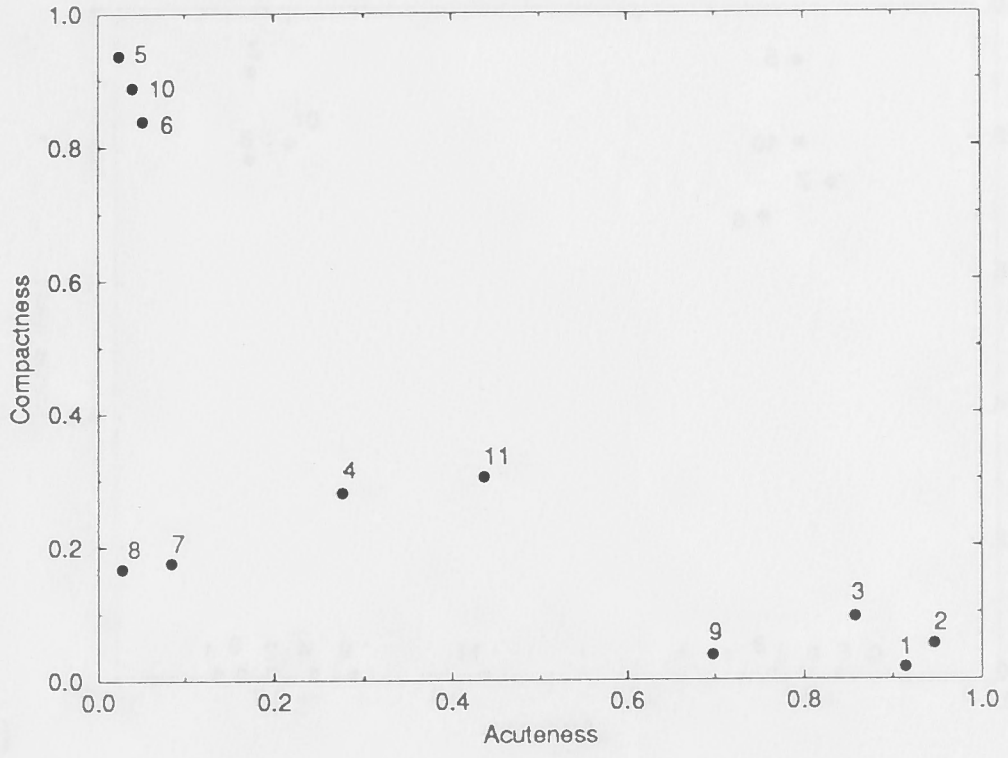


(b)

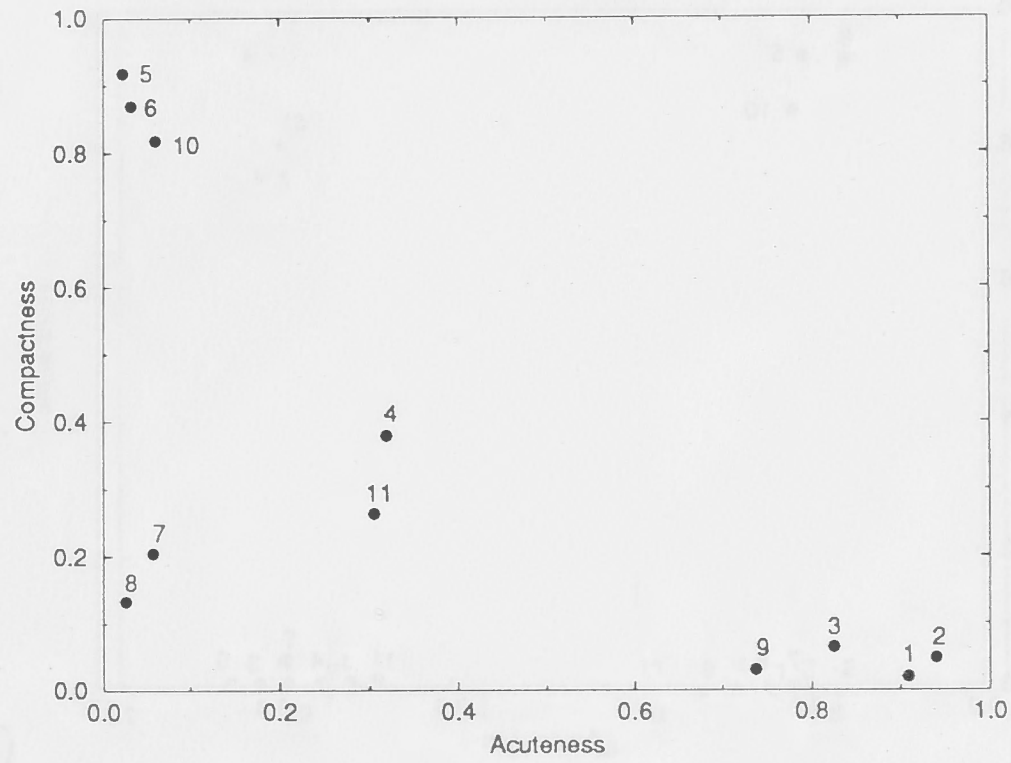


(c)

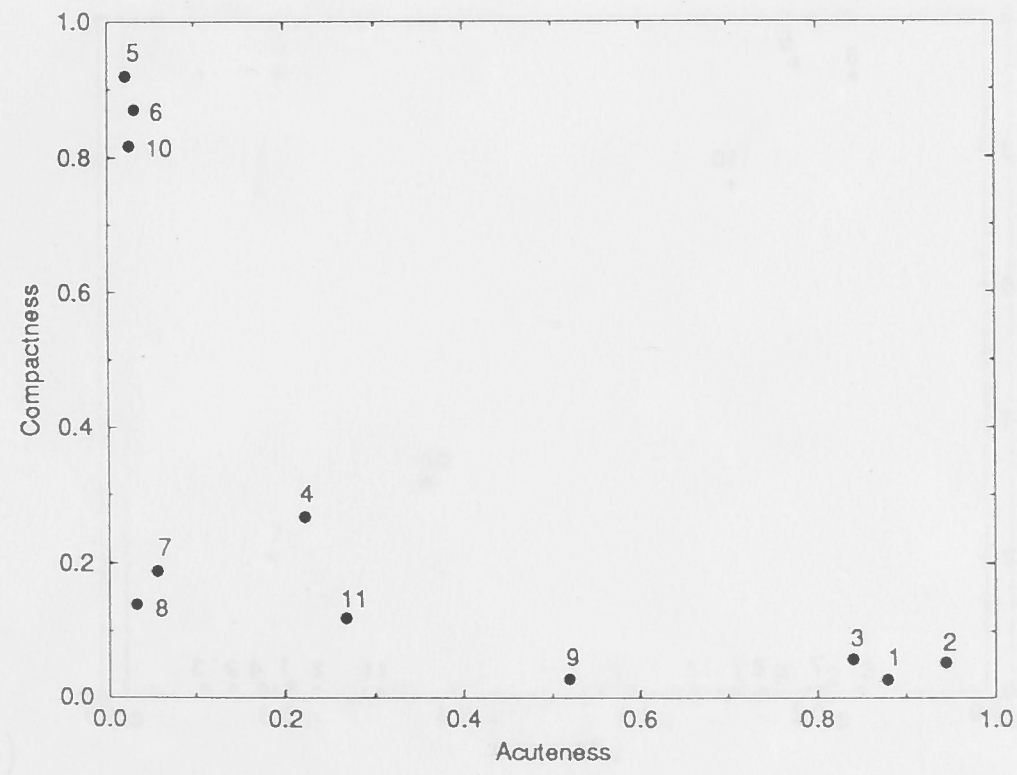
Figure A.50: *IM's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued on page 243*)



(d)

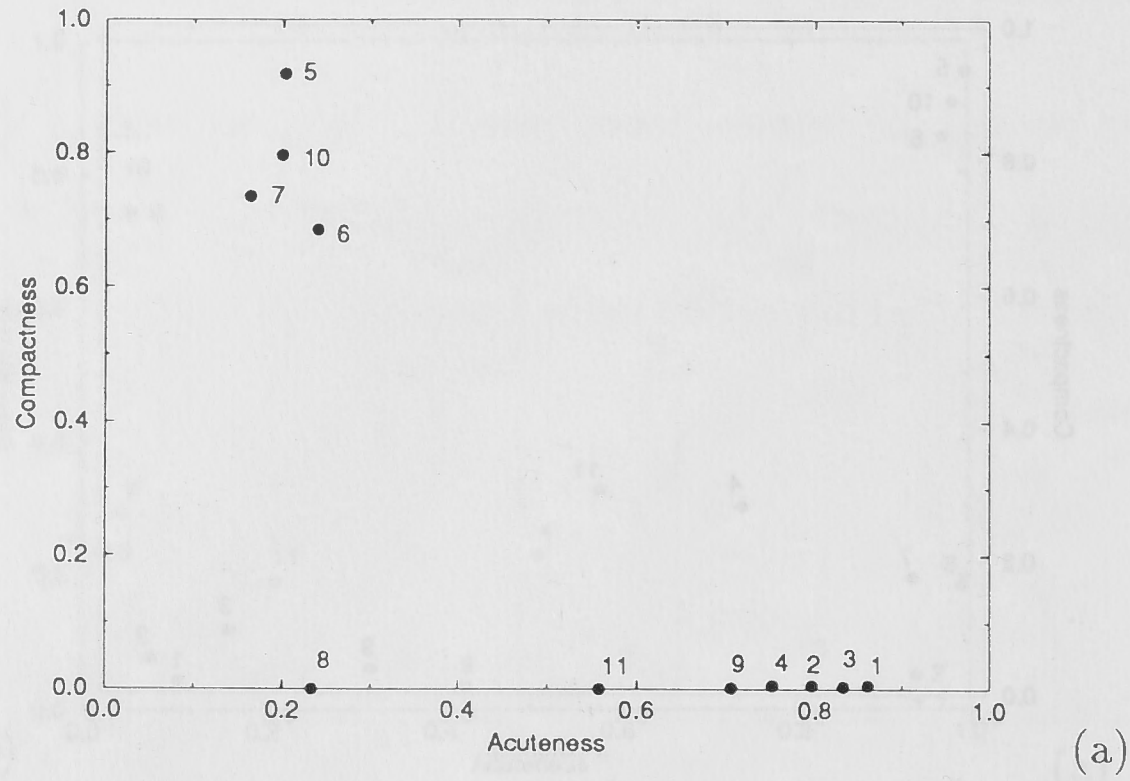


(e)

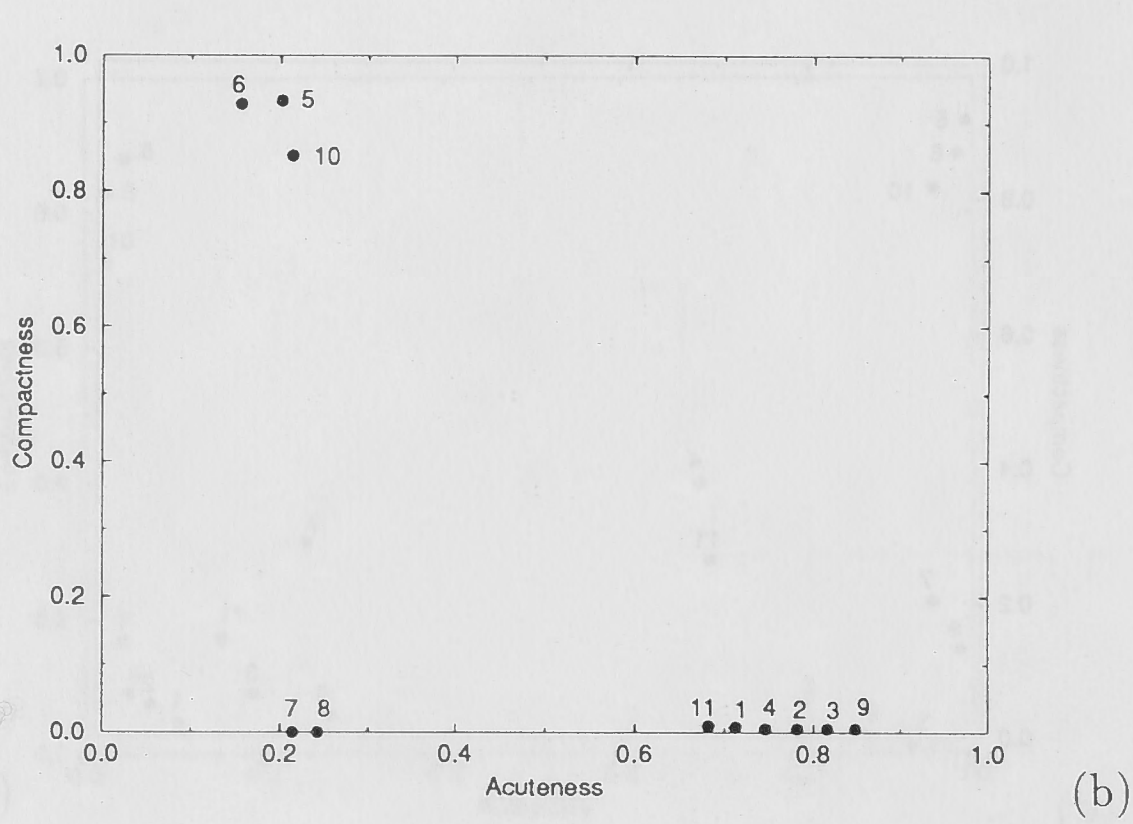


(f)

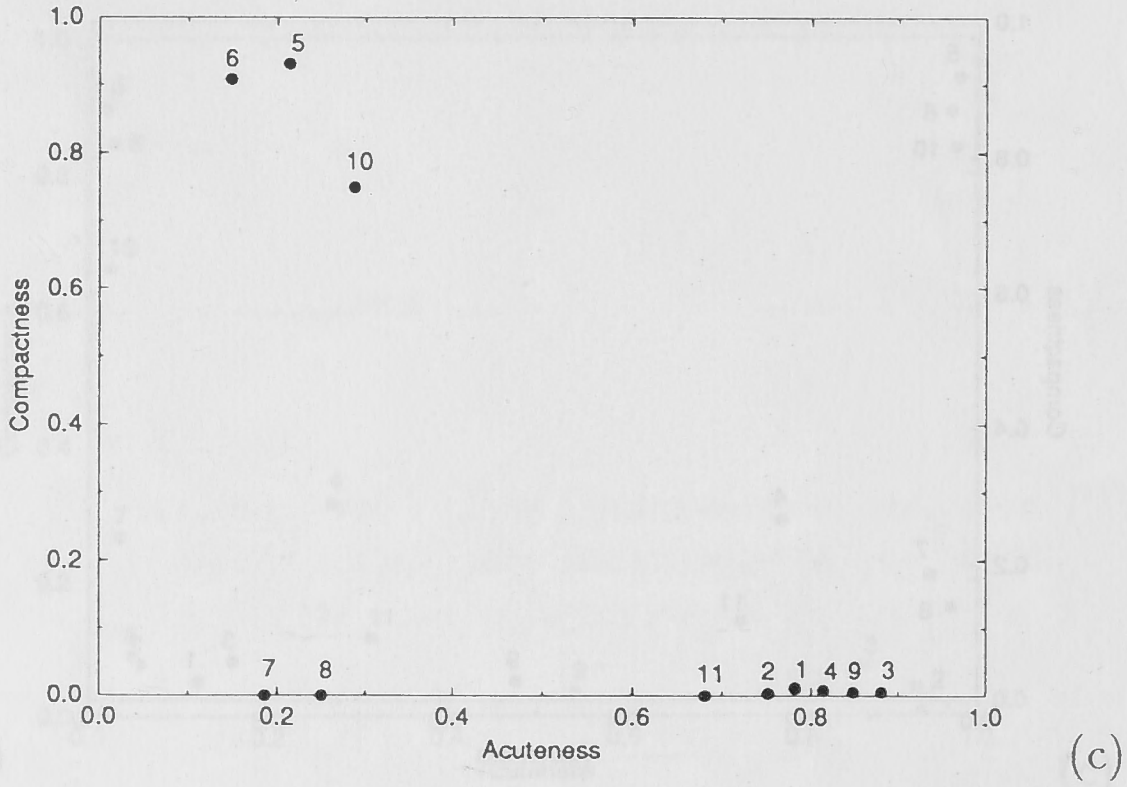
Figure A.50: *IM's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued from page 242*)



(a)

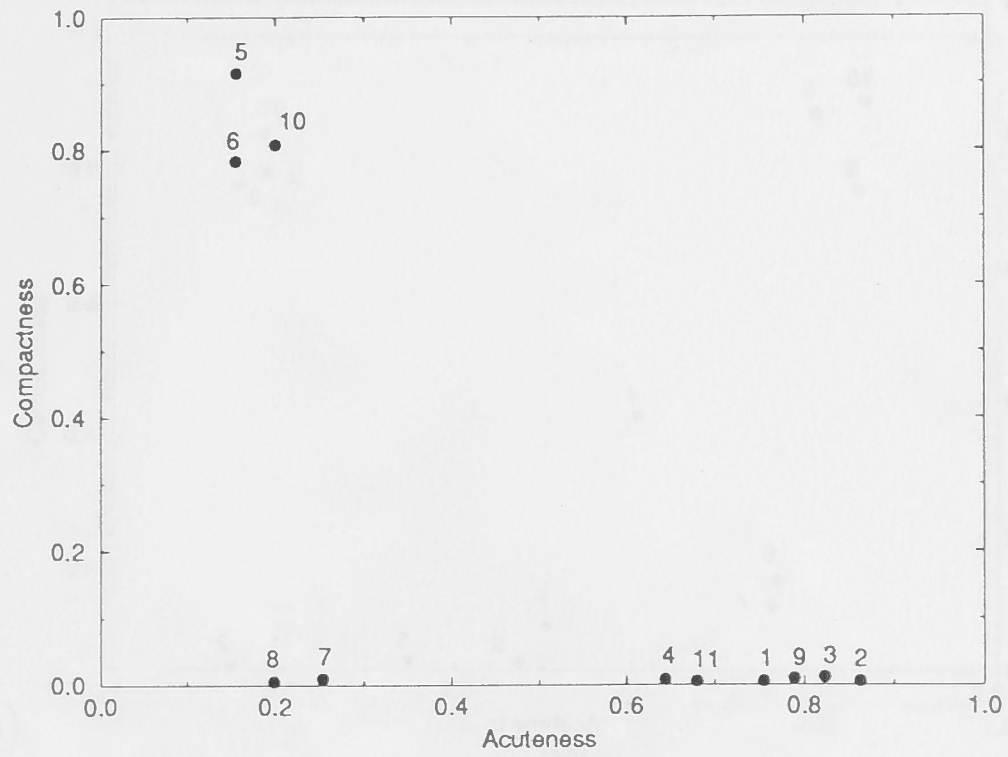


(b)

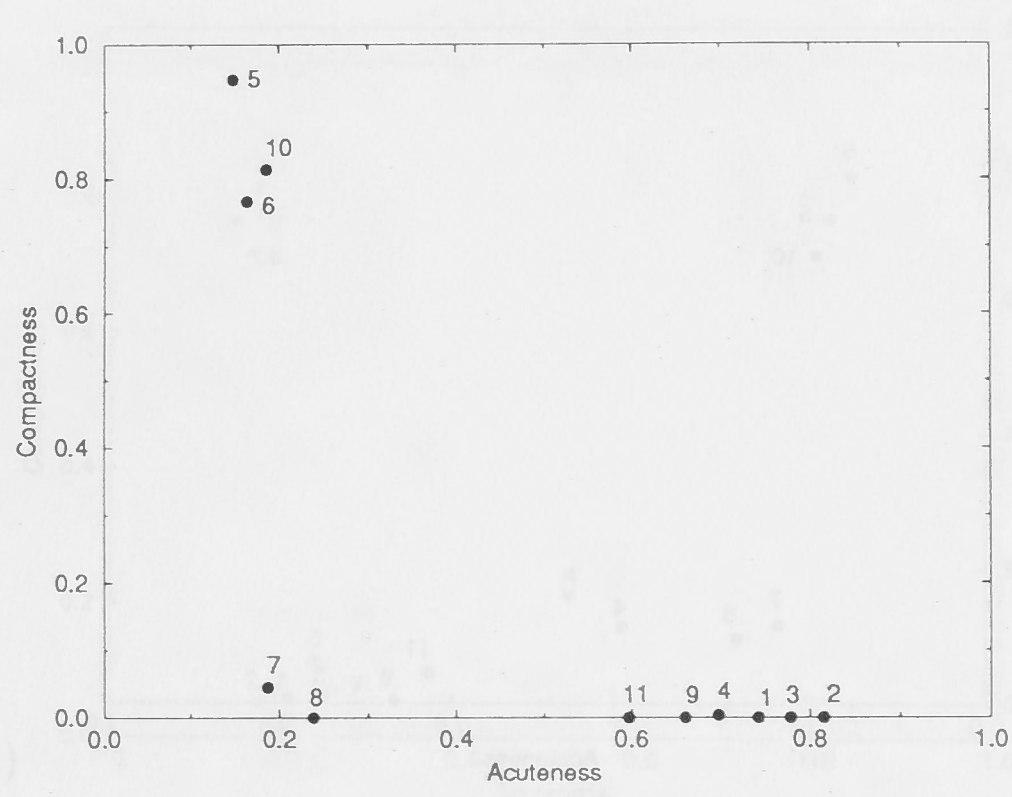


(c)

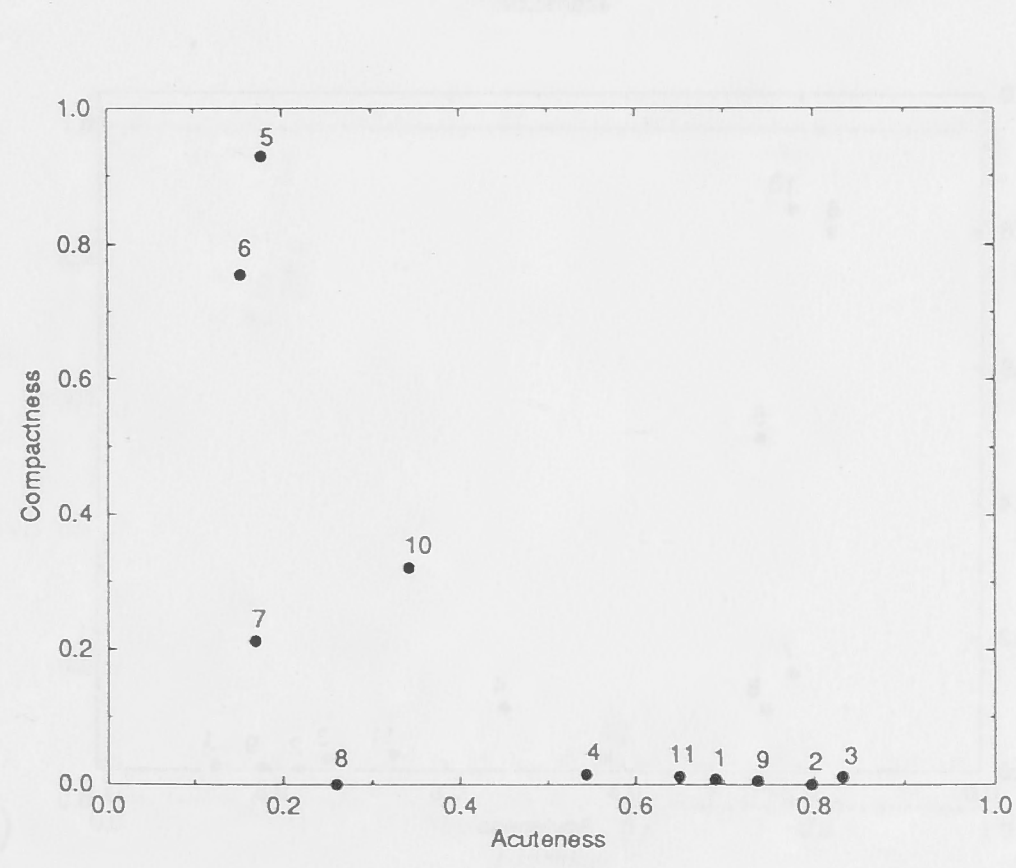
Figure A.51: *PM's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued on page 245*)



(d)

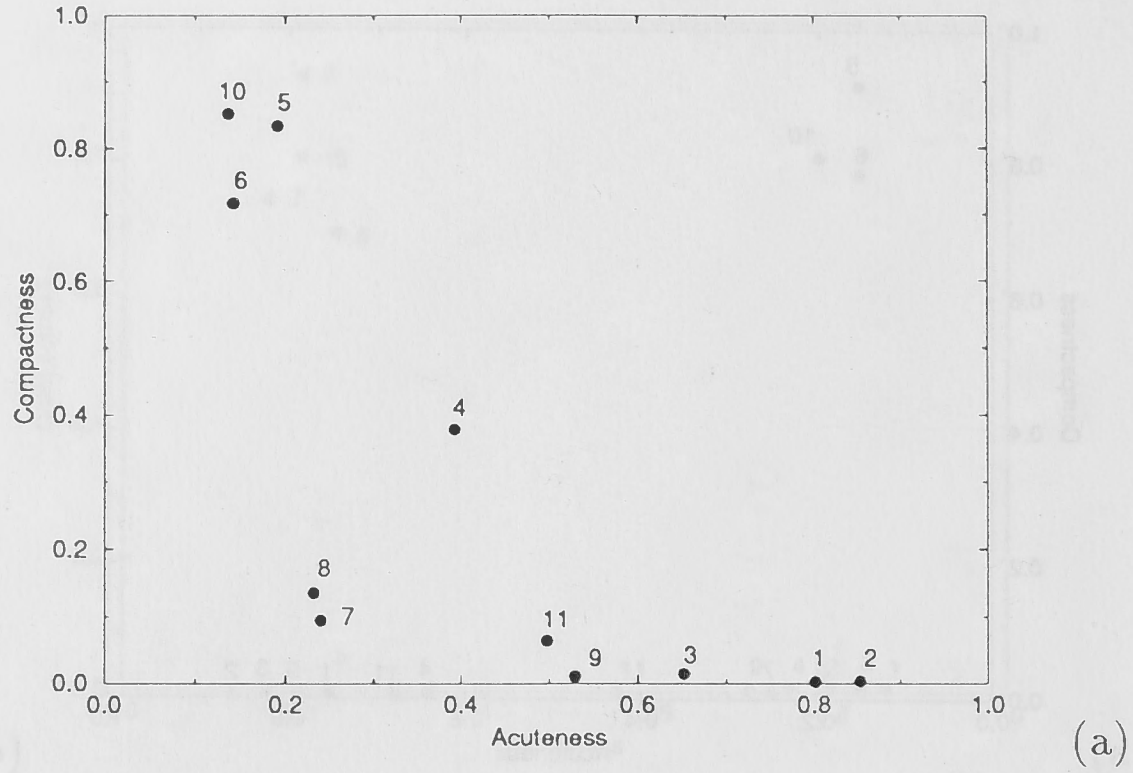


(e)

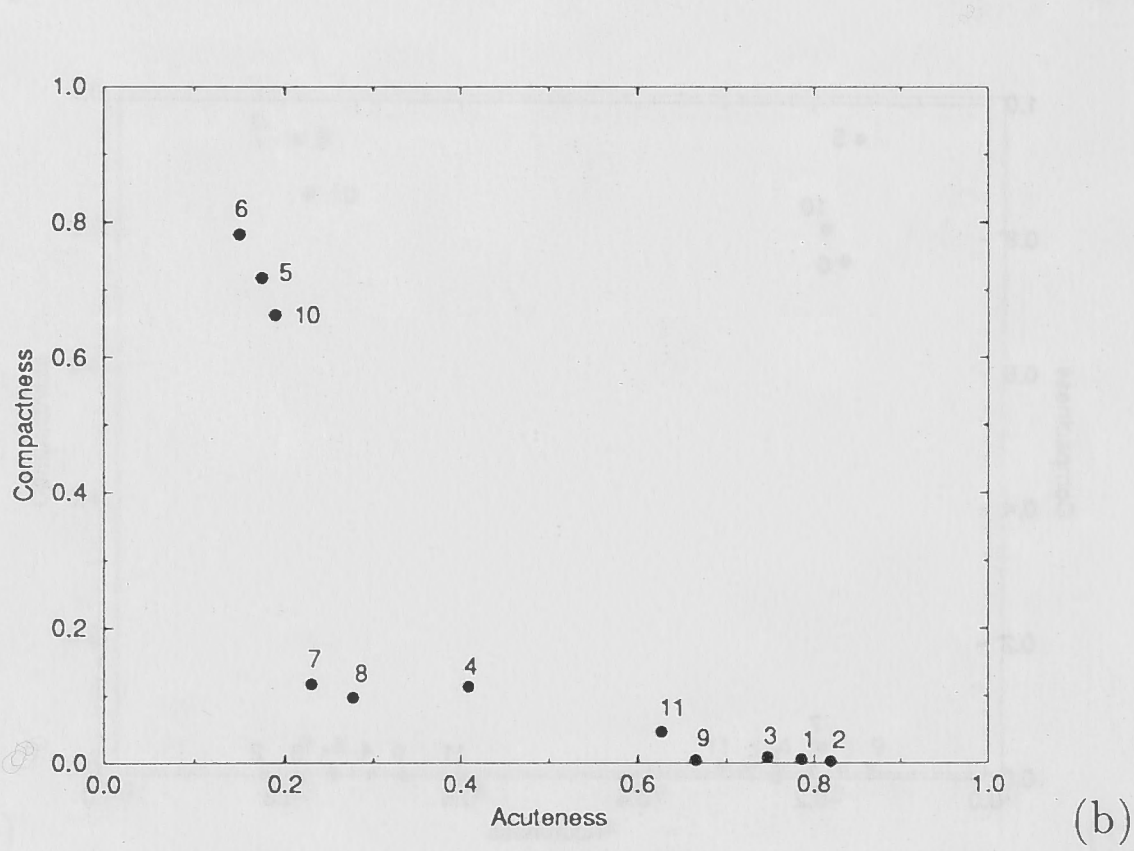


(f)

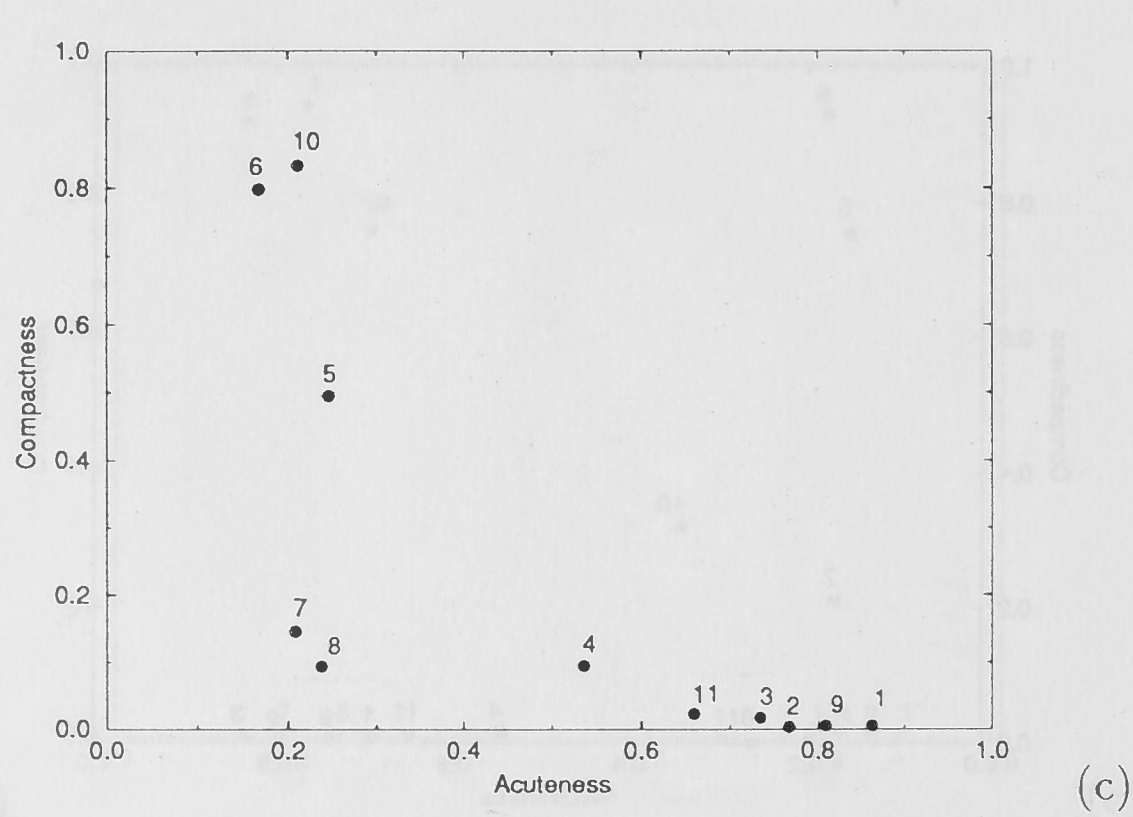
Figure A.51: *PM's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued from page 244*)



(a)

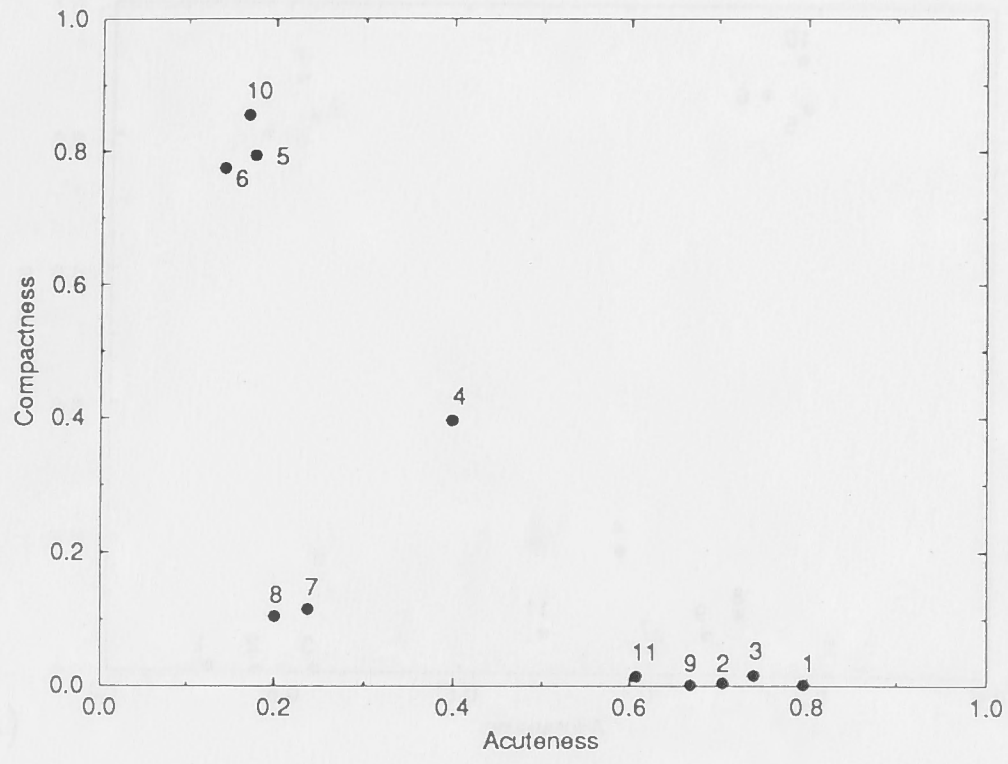


(b)

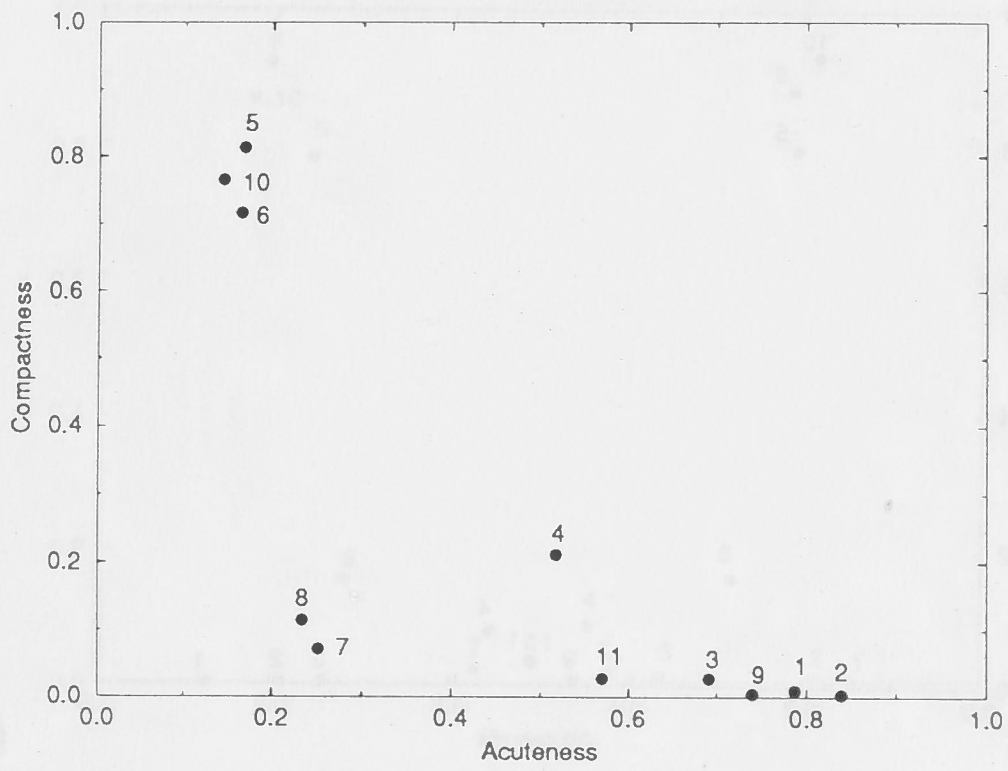


(c)

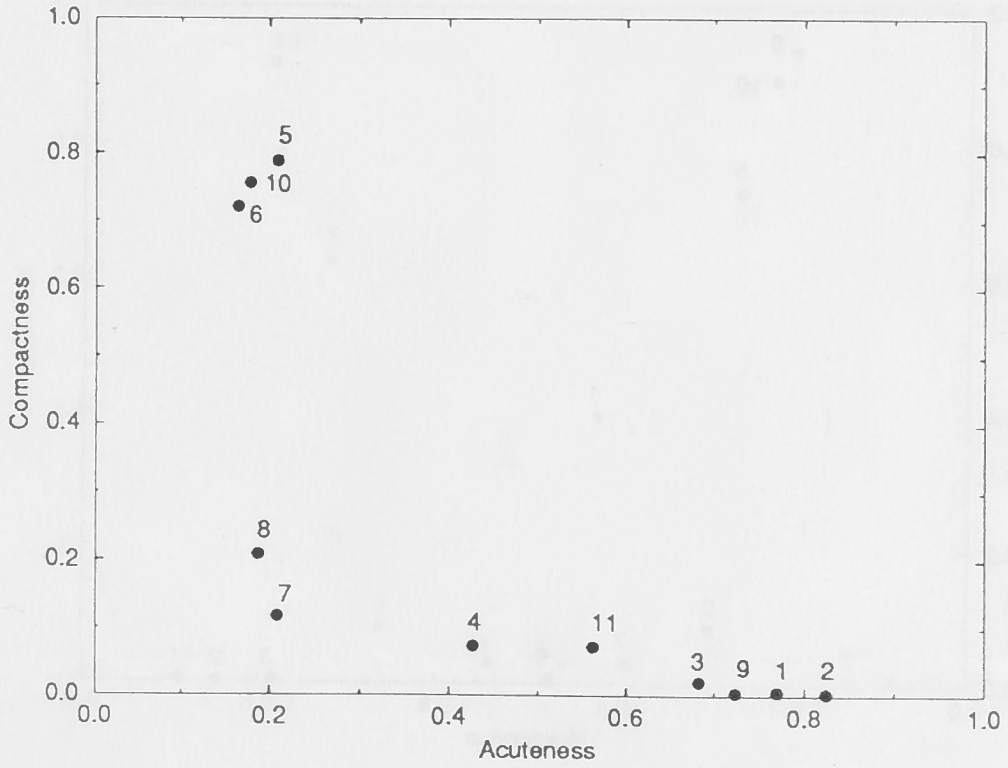
Figure A.52: *PP's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued on page 247*)



(d)

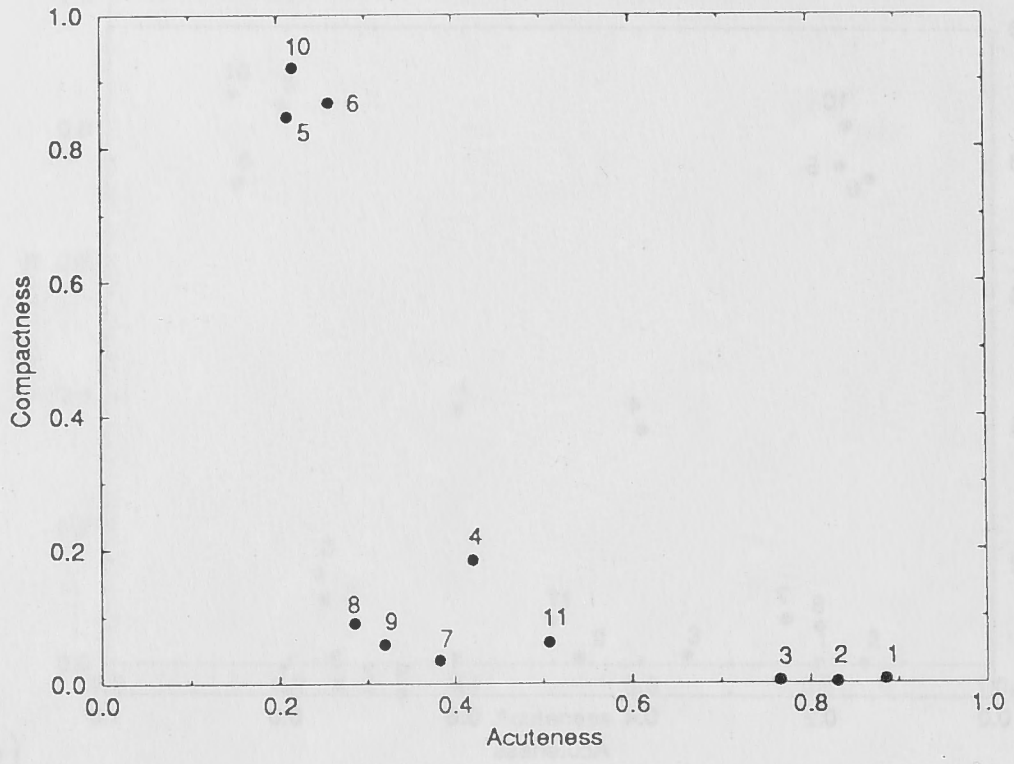


(e)

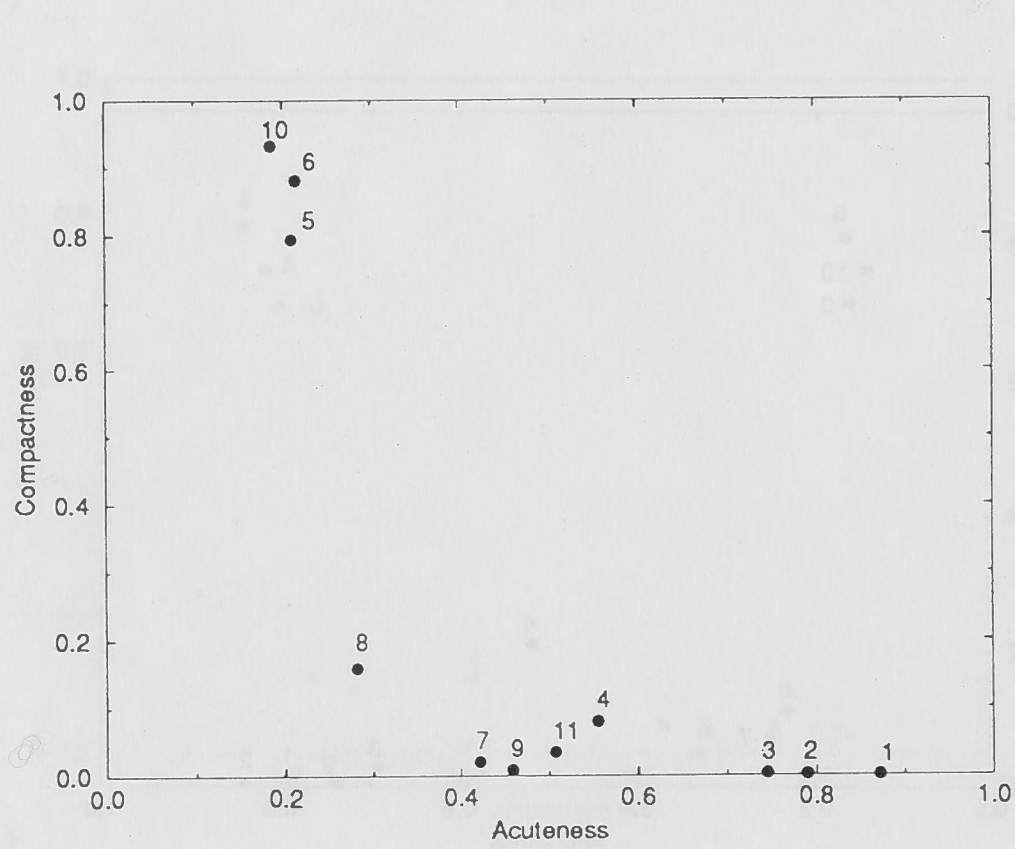


(f)

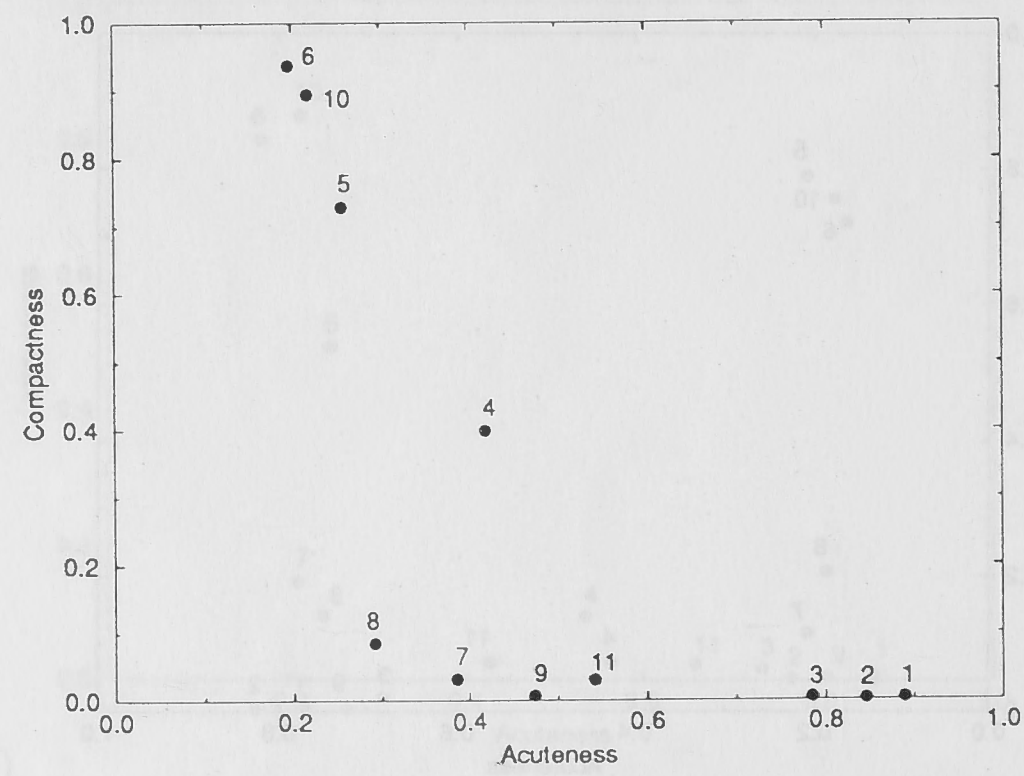
Figure A.52: *PP's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued from page 246*)



(a)

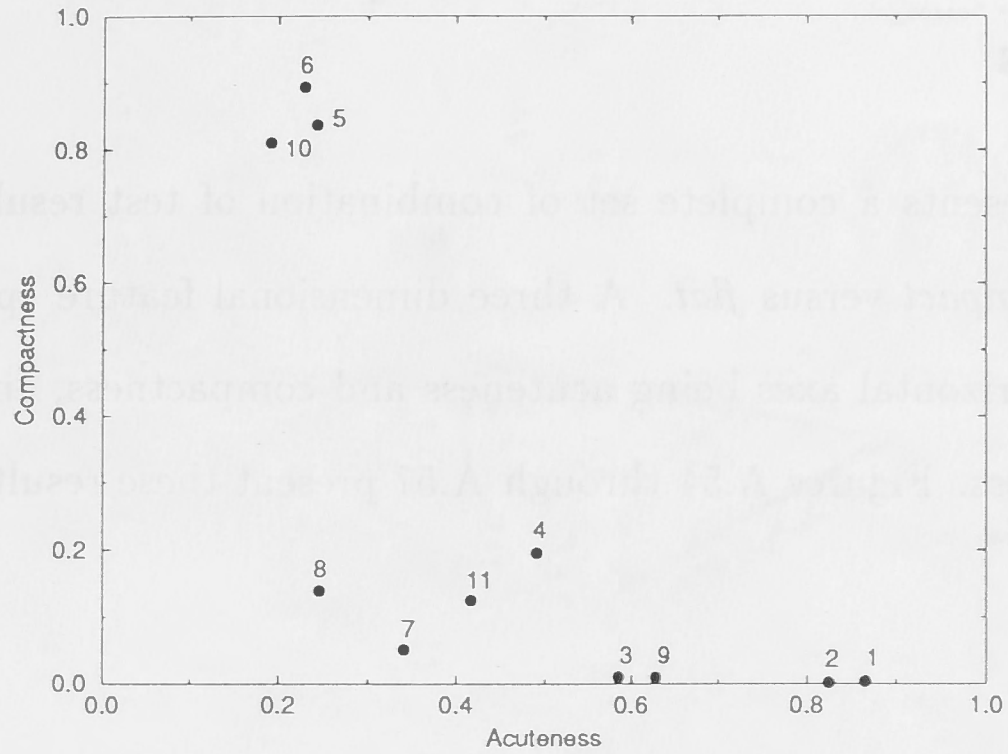


(b)

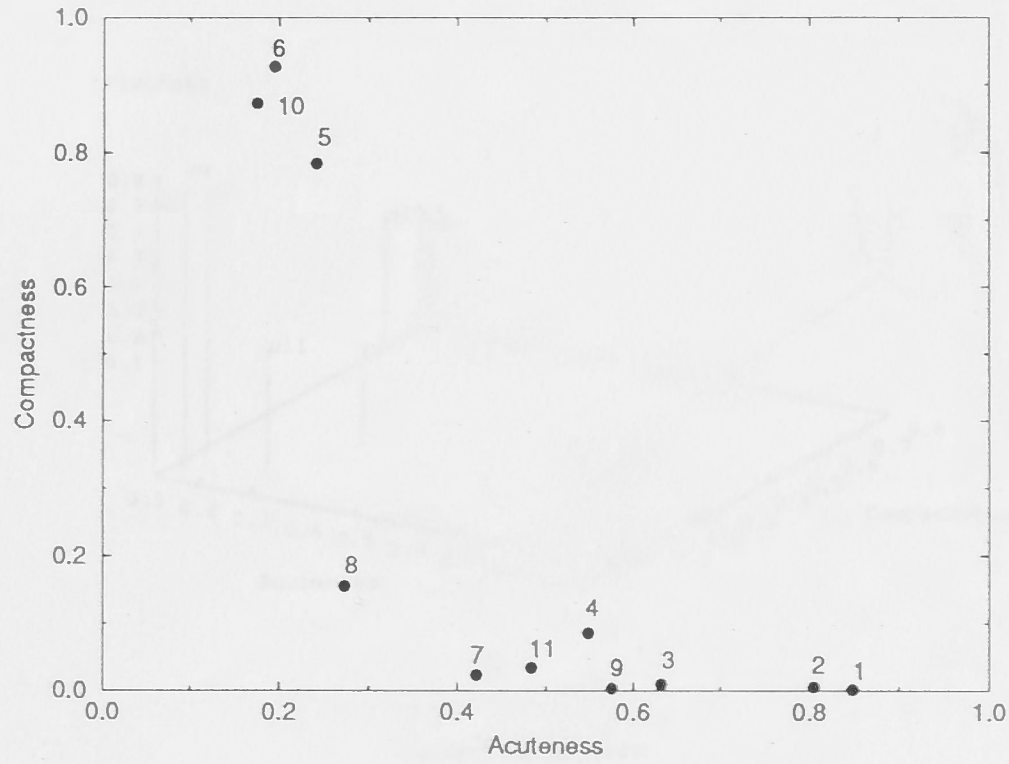


(c)

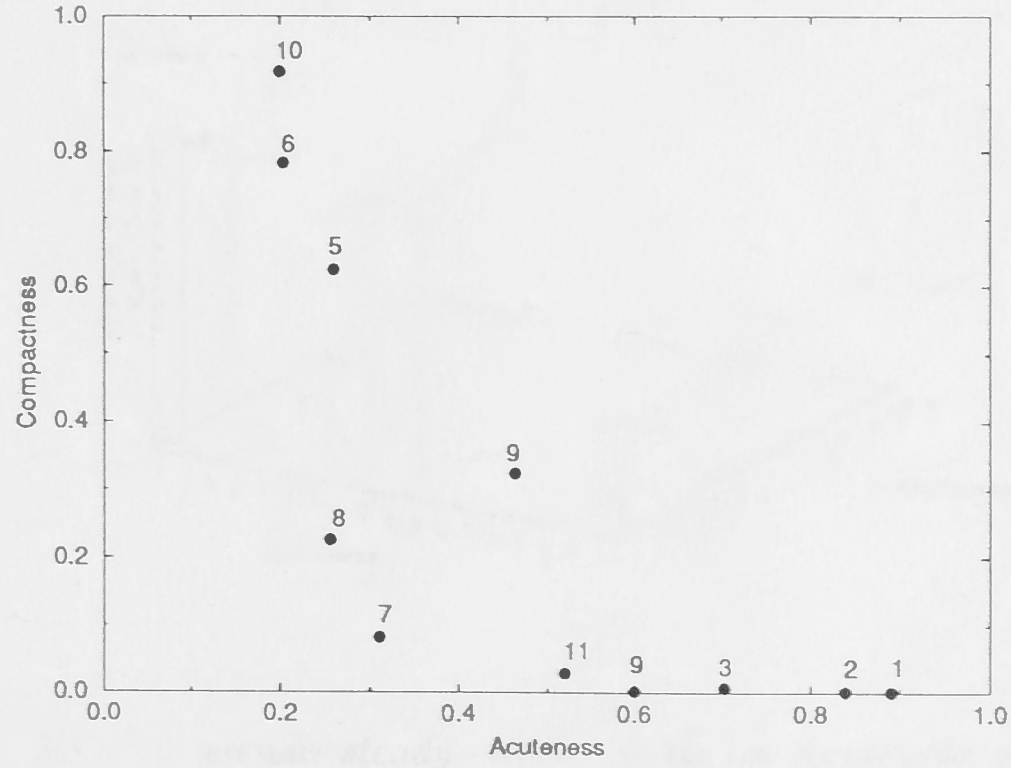
Figure A.53: *RP's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued on page 249*)



(d)



(e)

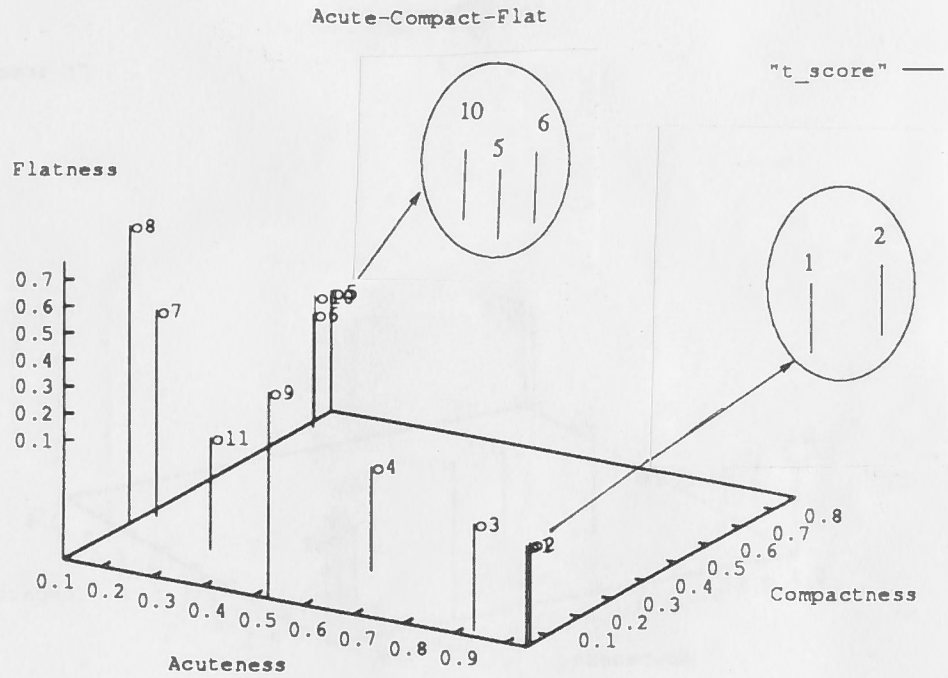


(f)

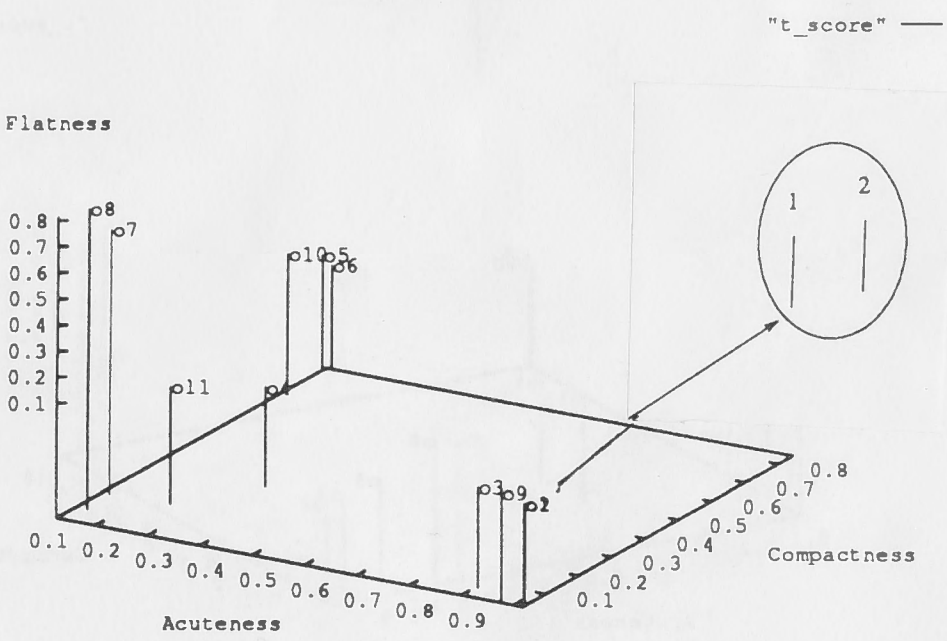
Figure A.53: *RP's 11 pseudo steady-state vowels on Acuteness versus Compactness plane in the context of six stop consonants:* (a) [b]; (b) [d]; (c) [g]; (d) [p]; (e) [t]; (f) [k]. Note order of the vowels: [i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ]. (*continued from page 248*)

A.8 Acuteness versus Compactness versus Flatness

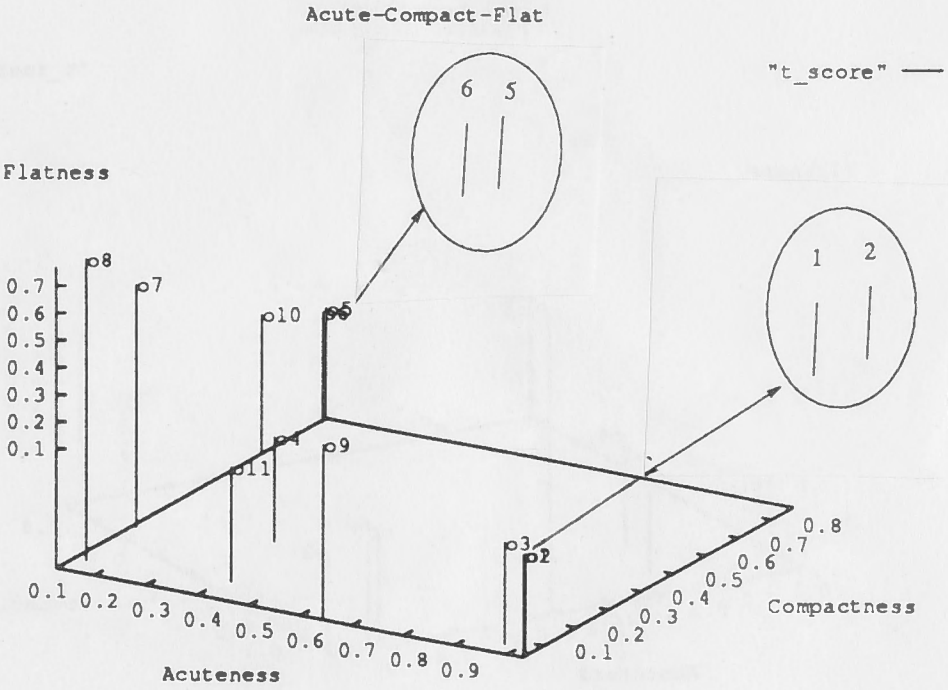
This section presents a complete set of combination of test results of features *acute* versus *compact* versus *flat*. A three dimensional feature space is formed with the two horizontal axes being acuteness and compactness, and the vertical axis being flatness. Figures A.54 through A.57 present these results.



(a)

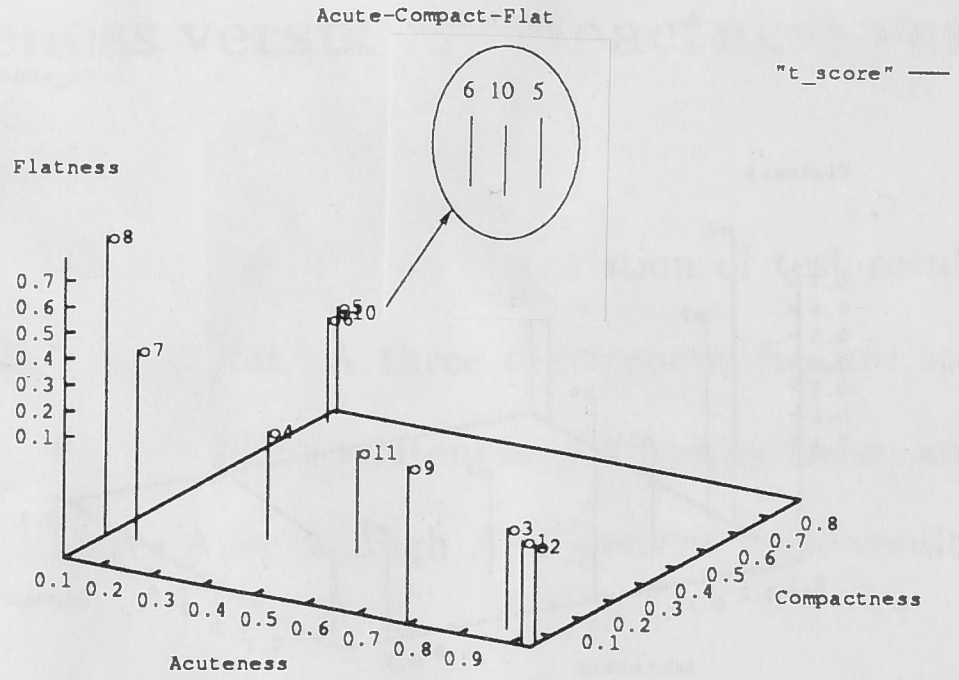


(b)

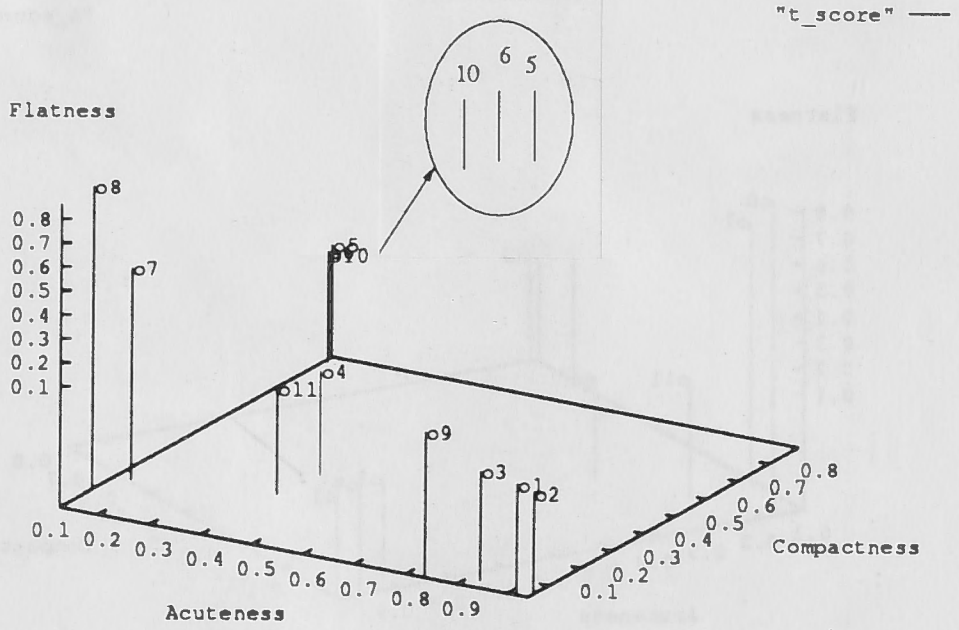


(c)

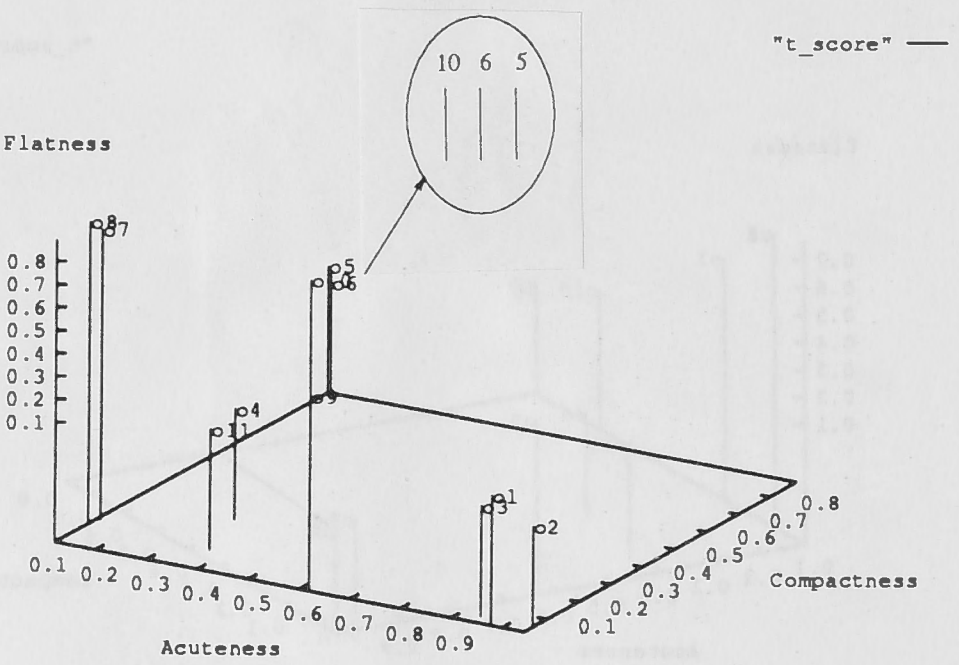
Figure A.54: *IM's 11 pseudo steady -state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued on page 252*)



(d)

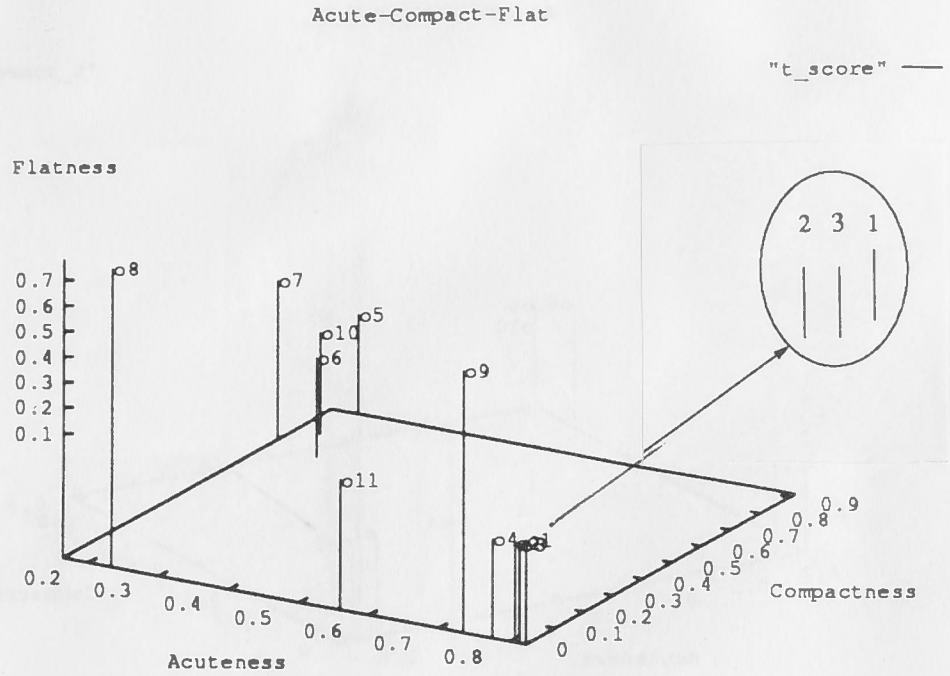


(e)

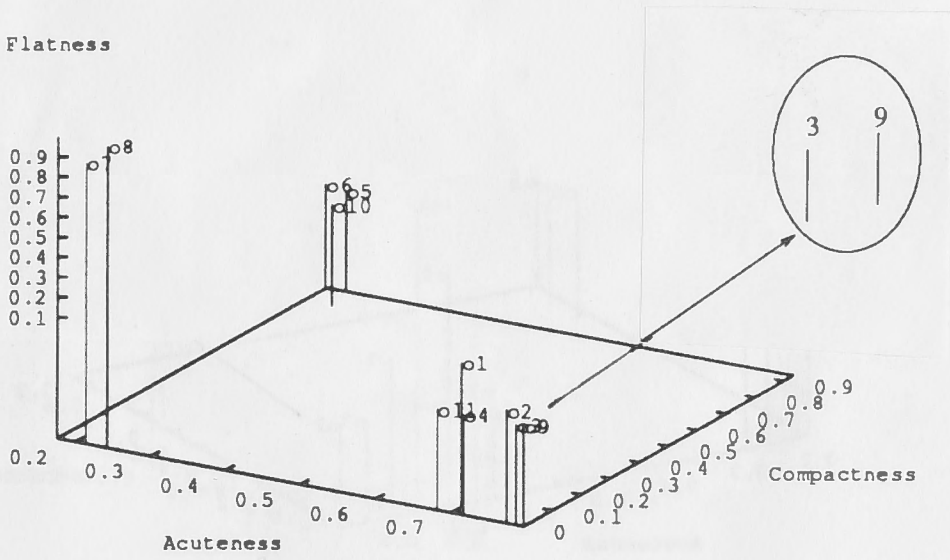


(f)

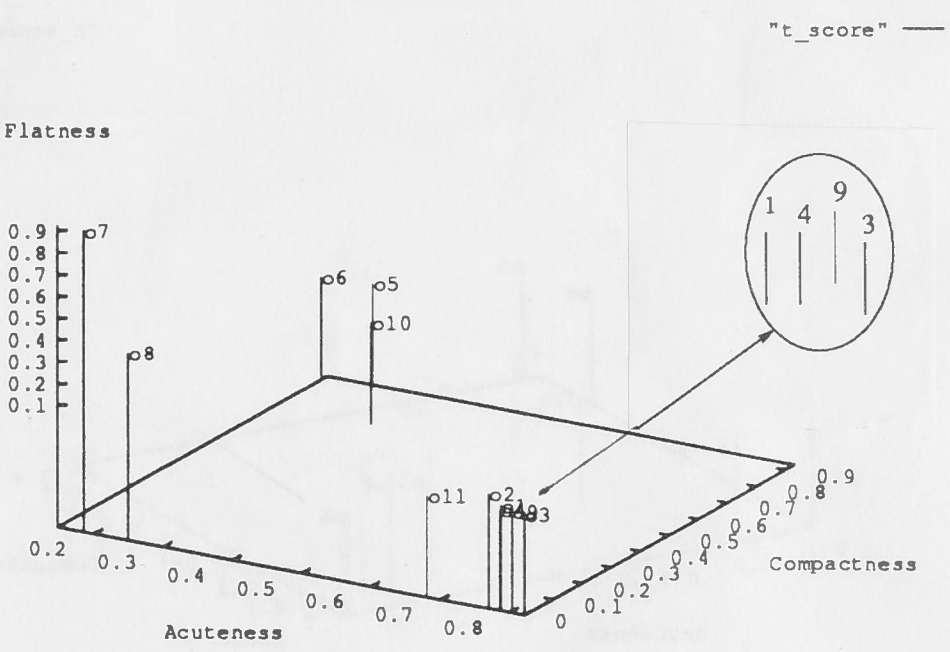
Figure A.54: *IM's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued from page 251*)



(a)

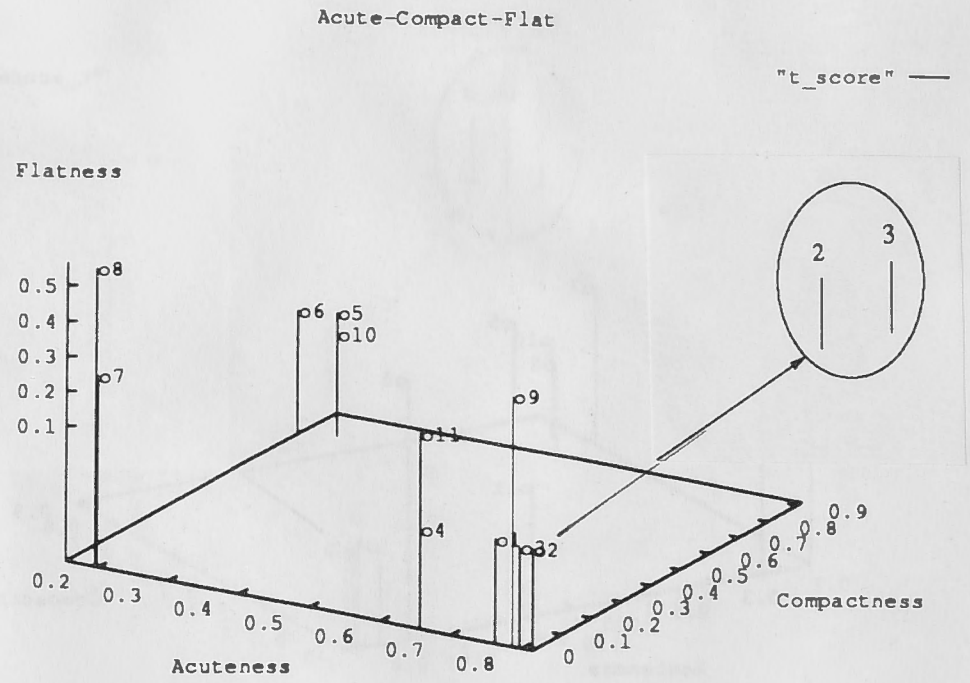


(b)

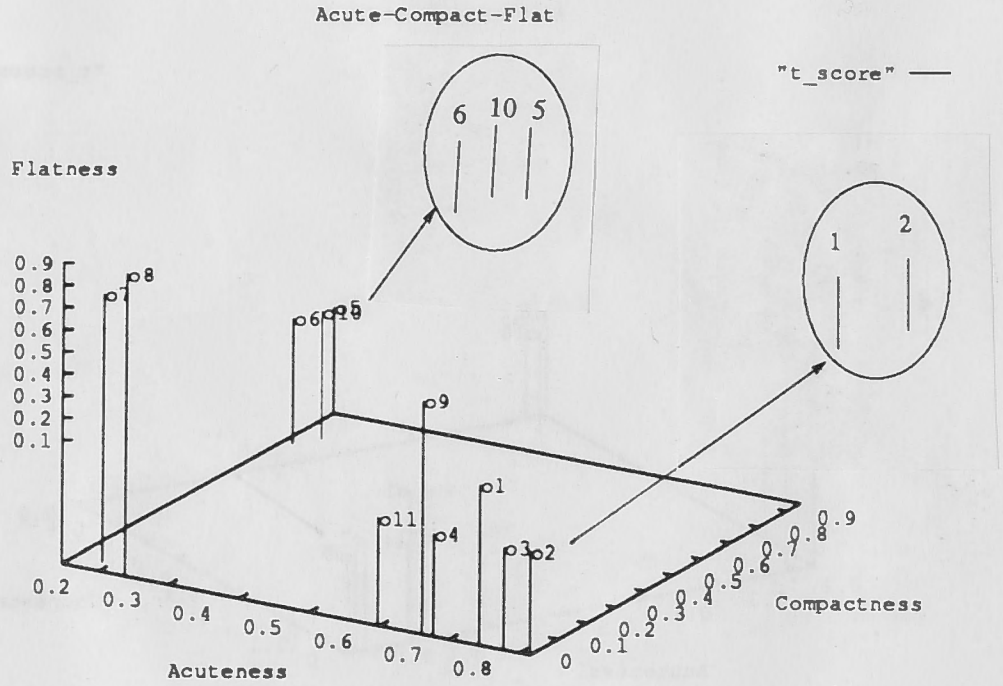


(c)

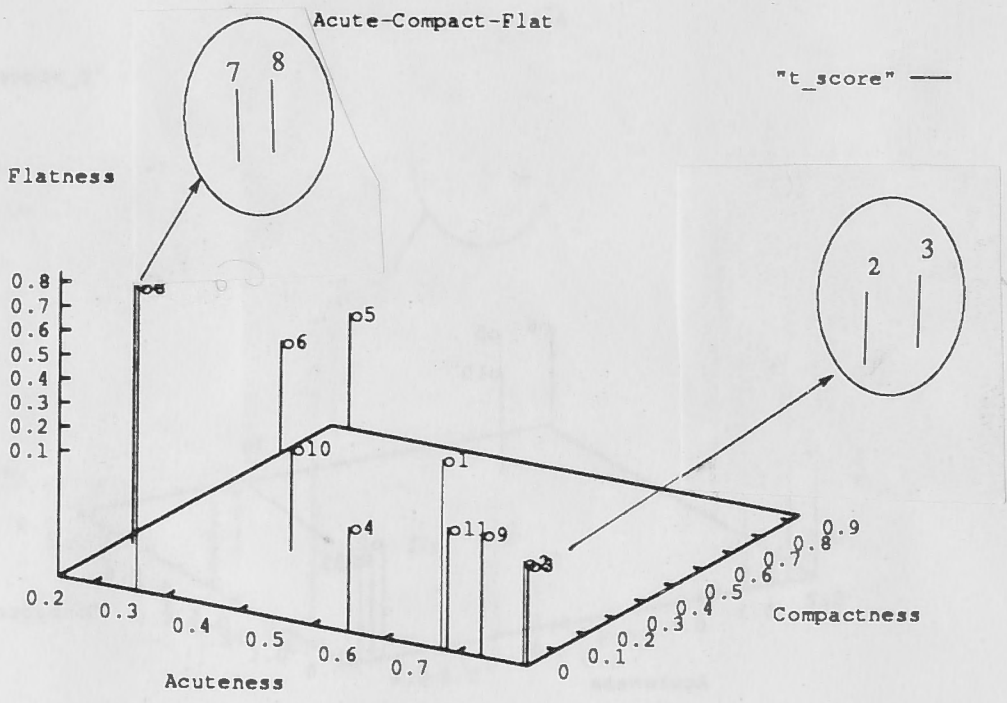
Figure A.55: *PM's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued on page 254*)



(d)

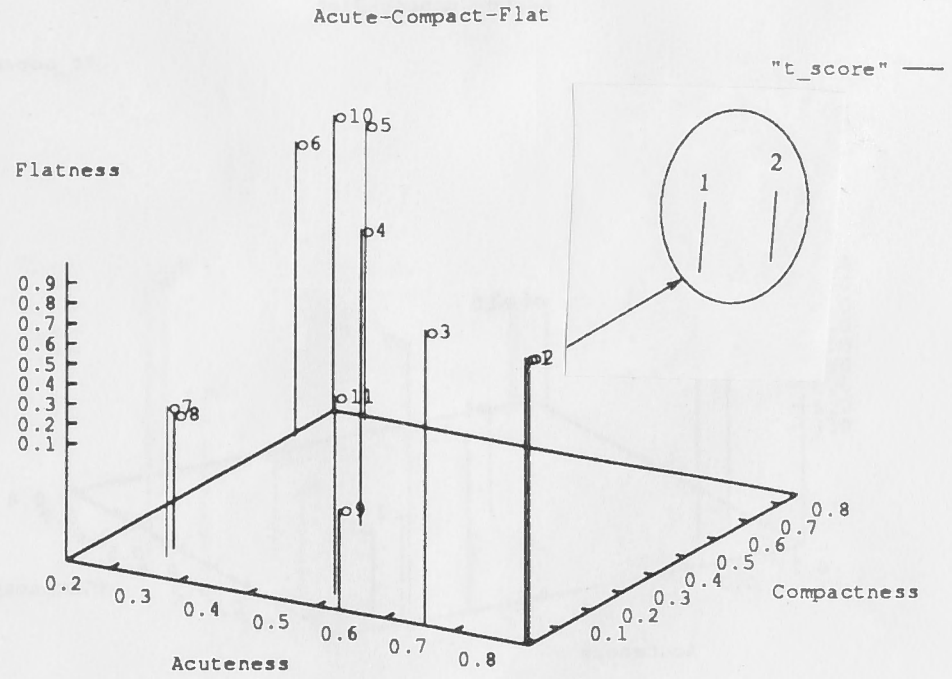


(e)

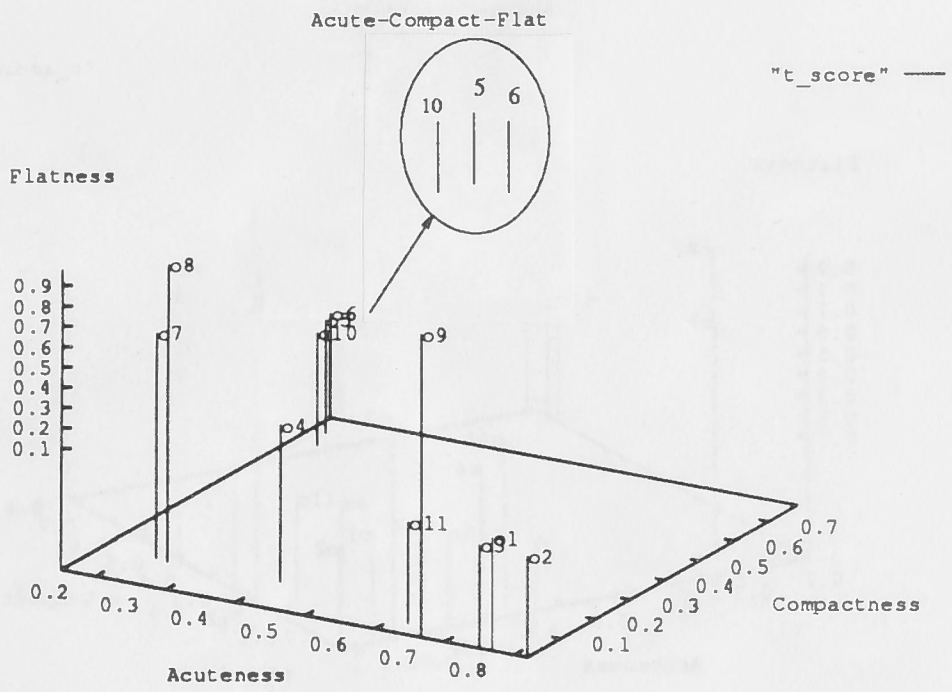


(f)

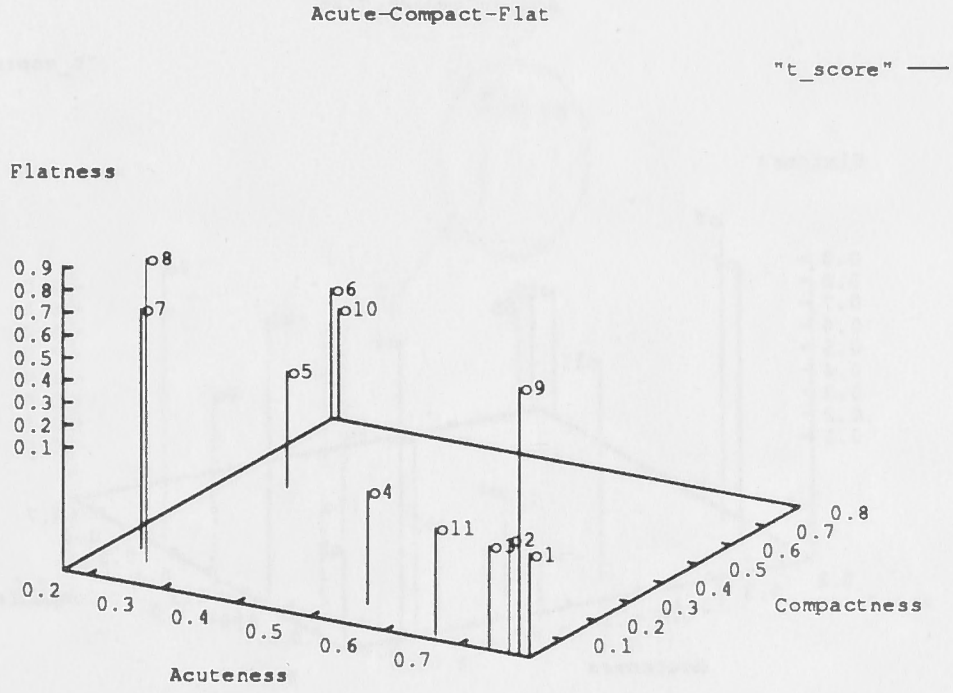
Figure A.55: *PM's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued from page 253*)



(a)

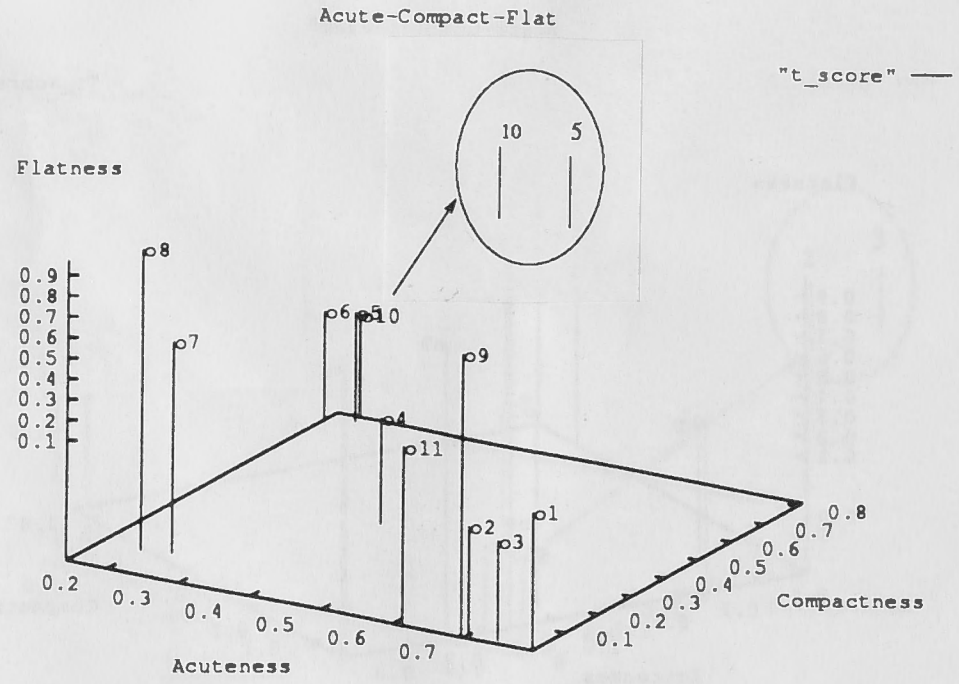


(b)

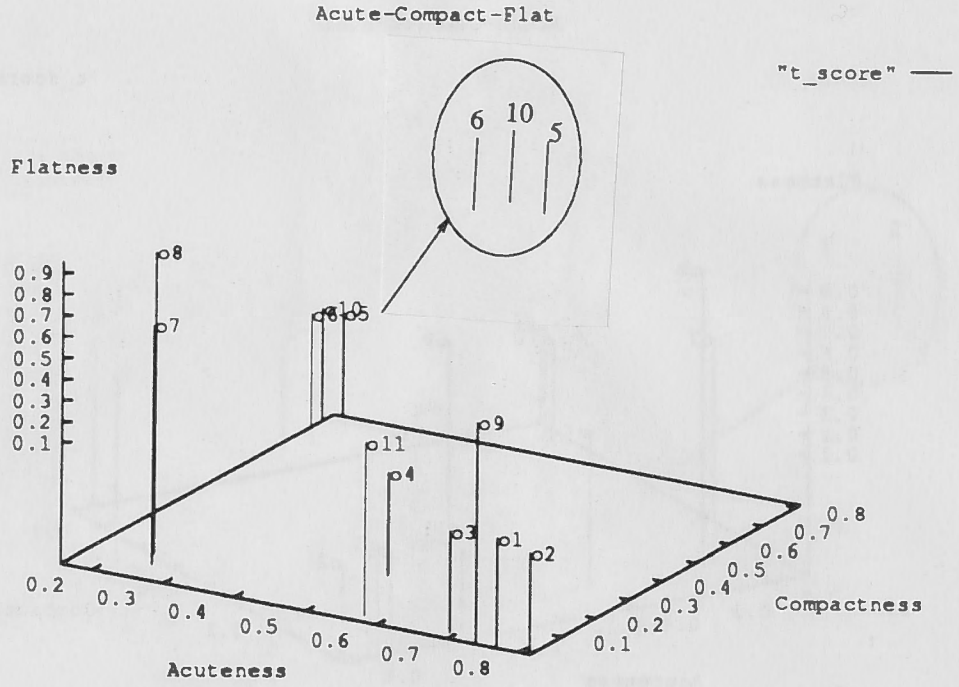


(c)

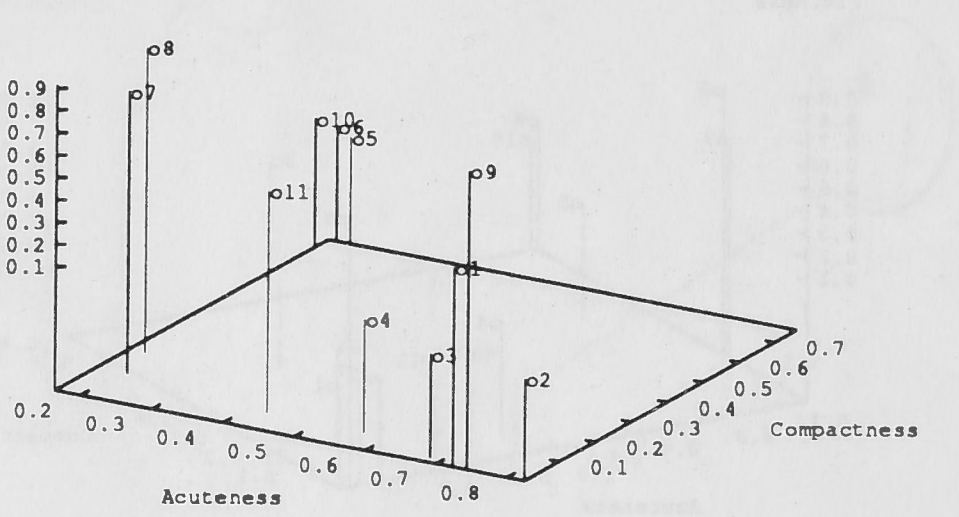
Figure A.56: *PP's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued on page 256*)



(d)

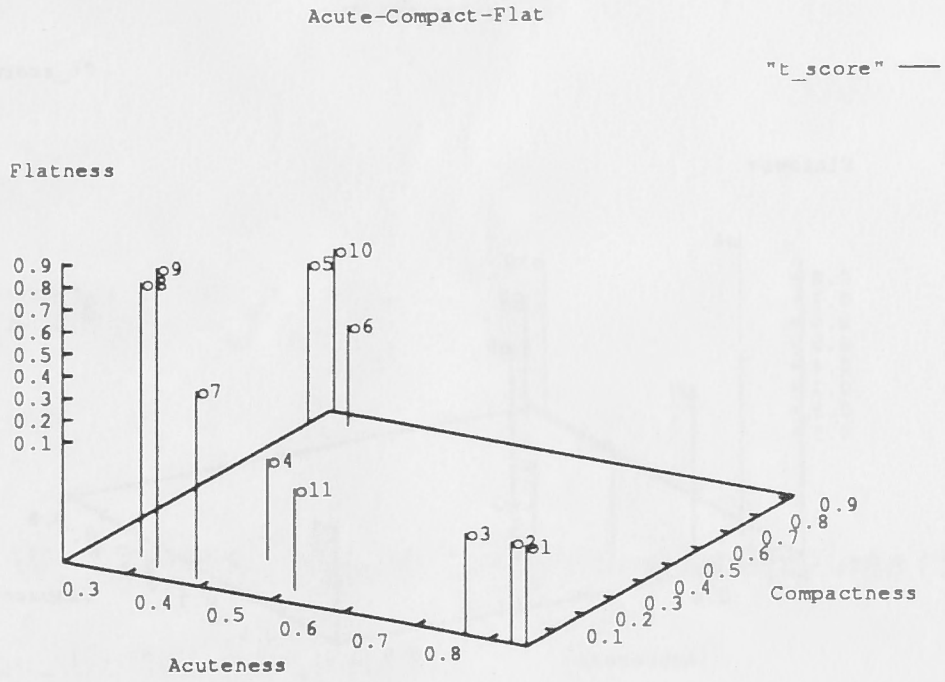


(e)

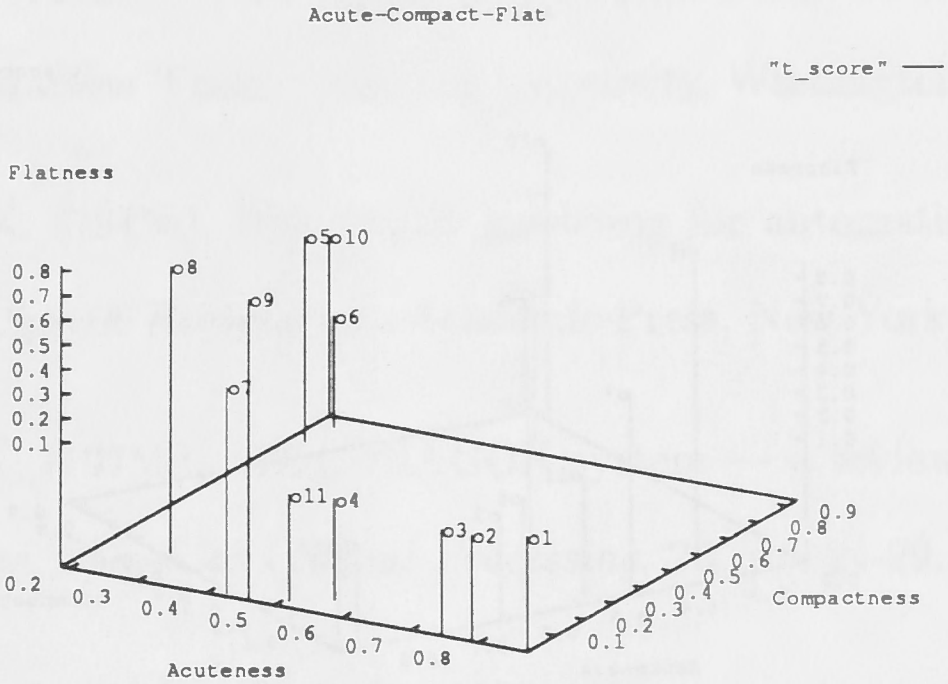


(f)

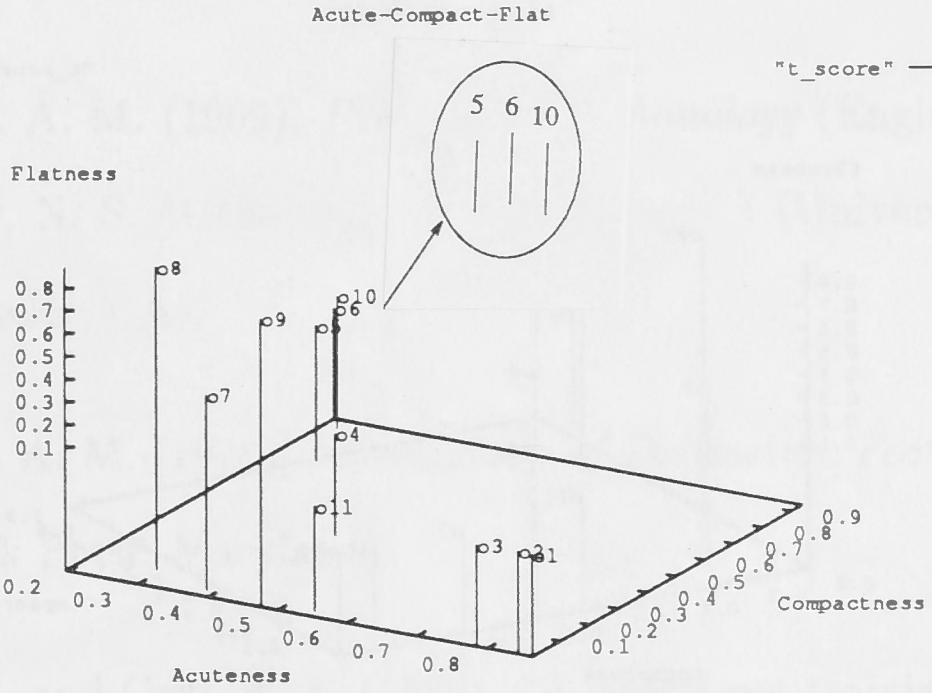
Figure A.56: *PP's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued from 255*)



(a)

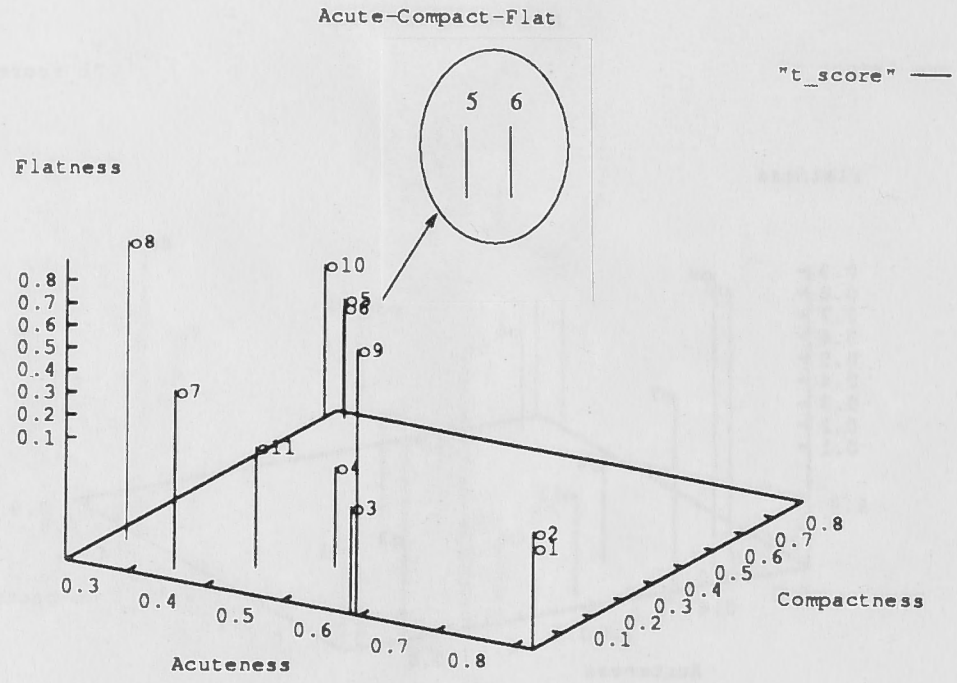


(b)

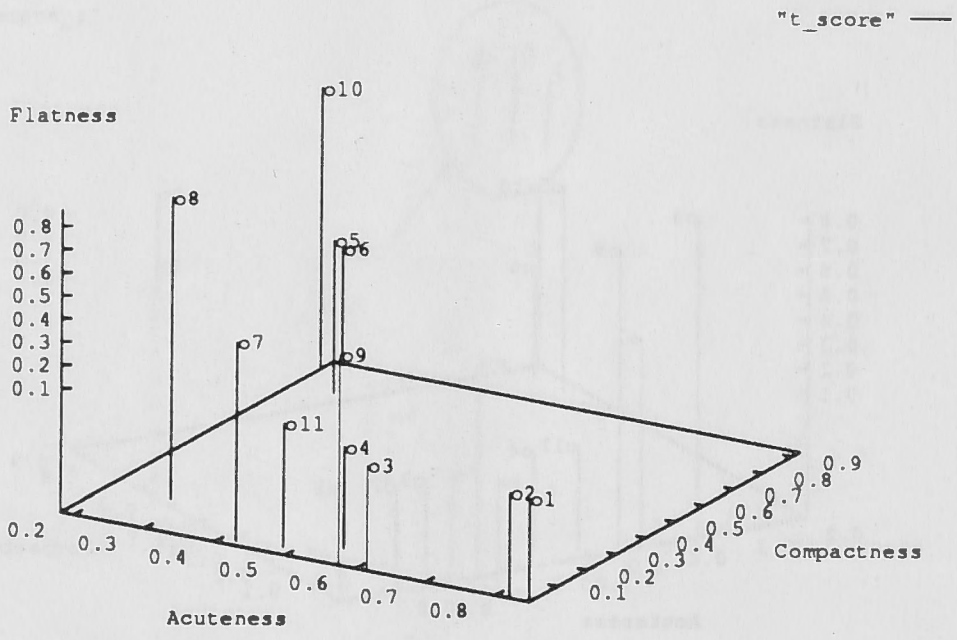


(c)

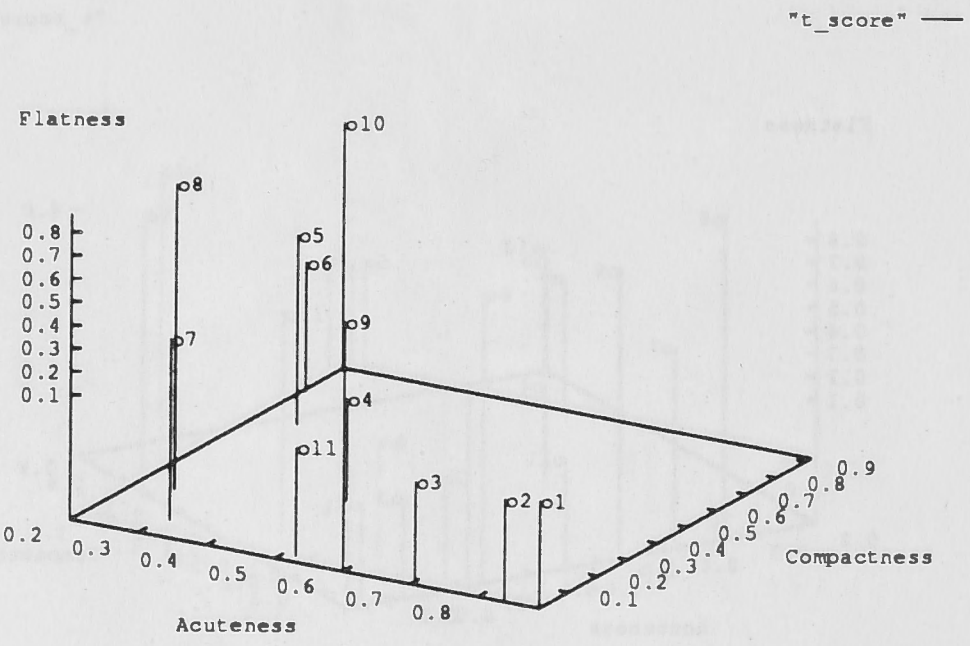
Figure A.57: *RP's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued on page 258*)



(d)



(a)



(f)

Figure A.57: *RP's 11 pseudo steady-state vowels on Acuteness versus Compactness versus Flatness plane in the context of six stop consonants:* (a) [**b**]; (b) [**d**]; (c) [**g**]; (d) [**p**]; (e) [**t**]; (f) [**k**]. Note order of the vowels: [**i**, **ɪ**, **ɛ**, **æ**, **ɑ**, **ɒ**, **ɔ**, **ʊ**, **u**, **ʌ**, **ɜ**]. (*continued from page 257*)

Bibliography

[1] Ainsworth, W. A. (1988), *Speech Recognition by Machine* (Peter Peregrinus Ltd., London, United Kingdom).

[2] Anglin, M. (1971), *Perceptual space of English vowels in word context*, Unpublished Masters Thesis, (Howard University, Washington, DC).

[3] Baker, J. K. (1975a), "Stochastic modeling for automatic speech understanding", *Speech Recognition* (Academic Press, New York), pp. 521-541.

[4] Baker, J. K. (1975b), "The DRAGON system — a review", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **23**, pp. 24-29.

[5] Balonov, L. J. and Deglin, V. L. (1976), *Slux i reč' dominantnogo i nedominantnogo polušarij* (Leningrad).

[6] Baltaxe, C. A. M. (1969), *Principles of Phonology* (English translation of Trubetzkoy, N. S. "Grundzüge der phonologie") (University of California Press, Berkeley, CA).

[7] Baltaxe, C. A. M. (1978), *Foundations of Distinctive Feature Theory* (University Park Press, Maryland).

[8] Barnard, E. and Cole, R. A. (1989), "A neural-net training program based on conjugate-gradient optimization", *Technical Report No. CSE 89-014* (Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA).

[9] Barr, A. and Feigenbaum, E. A. (1982), *The Handbook of Artificial Intelligence* (HeirisTech Press, Stanford, CA).

[10] Baum, L. E. (1972), "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes", *Inequalities*, **3**, pp. 1-8.

[11] Bellegarda, J. R. and Nahamoo, D. (1989), "Tied mixture continuous parameter models for large vocabulary isolated speech recognition", *Proc. IEEE ICASSP*, pp. 13-16.

[12] Bellman, R. (1957), *Dynamic Programming* (Princeton University Press, Princeton, NJ).

[13] Bernard, J. R. (1989), "Quantitative aspects of the sounds of Australian English", in Collins, P. and Blair, D. (Eds.), *Australian English: The Language of A New Society* (University of Queensland Press, Australia), pp. 87-204,

[14] Bimbot, F., Chollet, G. and Tubach, J-P. (1991), "TDNNs for phonetic features extraction: a visual exploration", *Proc. IEEE ICASSP*, pp. 73-76.

[15] Blache, S. E. (1978), *The Acquisition of Distinctive Features* (University Park Press, Maryland).

[16] Bloomfield, L. (1933), *Language* (George Allen & Unwin Ltd., New York).

[17] Blumstein, S. E. and Cooper, W. (1972), "Identification versus discrimination of distinctive features in speech perception", *Quarterly Journal of Experimental Psychology*, **24**, pp. 207-214.

[18] Blumstein, S. E. (1974), "The use and theoretical implications of dichotic technique for investigating distinctive features", *Brain and Language*, **1**, pp. 337-350.

[19] Blumstein, S. E. and Stevens, K. N. (1979), "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *J. Acoust. Soc. Am.*, **66**, pp. 1001-1017.

[20] Blumstein, S. E. (1986), "On acoustic invariance in speech", in Perkell, J. S. and Klatt, D. M. (Eds.), *Variability and Invariance in Speech Processes* (Lawrence Erlbaum Assoc., Hillsdale, NJ), pp. 178-193.

[21] Bourlard, H. and Wellekens, C. J. (1987), "Speech pattern discrimination and multilayer perceptrons", *Research Report M.211* (Philips Research Laboratory).

[22] Bourlard, H. and Morgan, N. (1990), "A continuous speech recognition system embedding MLP into HMM", in Touretzky, D. (Ed.), *Advances in Neural Information Processing Systems 2* (Morgan Kaufmann, Los Altos, CA), pp. 186-193.

[23] Bourlard, H. and Wellekens, C. J. (1990), "Links between Markov models and multilayer perceptrons", *IEEE Trans. Pattern Anal. Machine Intell.*, **12**(12), pp. 1167-1178.

[24] Bourlard, H., Morgan, N., and Renals, S. (1992), "Neural nets and hidden Markov models: Review and generalizations", *Speech Communication*, **11**, pp. 237-246.

[25] Brakel, A. (1983), *Phonological Markedness and Distinctive Features* (Indiana University Press, Bloomington, IN).

[26] Bridle, J. S. and Brown, M. D. (1979), "Connected word recognition using whole word templates", *Proc. Acoust. Autumn Conf.*, pp. 25-28.

[27] Bridle, J. S., Brown, M. D. and Chamberlain, R. M. (1982), "An algorithm for connected word recognition", *Proc. IEEE ICASSP*, pp. 899-902.

[28] Chester, D. L. (1990), "Why two hidden layers are better than one", *Proc. Intl Joint Conf. on Neural Networks*, **1**, pp. 265-268.

[29] Chistovich, L., Fant, G., de Serpa-Leitão, A. and Tjernlund, P. (1966a), "Mimicking of synthetic vowels", *Speech Transmission Lab. Quart. Progr. Status Report* (Royal Inst. Tech. Stockholm), **2**, pp. 1-18.

[30] Chistovich, L., Fant, G., and de Serpa-Leitão, A. (1966b), "Mimicking and perception of synthetic vowels, part II", *Speech Transmission Lab. Quart. Progr. Status Report* (Royal Inst. Tech. Stockholm), **3**, pp. 1-3.

[31] Chomsky, N. and Halle, M. (1968), *The Sound Pattern of English* (Harper & Row, Publishers, New York).

[32] Clermont, F. and Millar, B. (1986), "Multi-speaker validation of coarticulation models of syllabic nuclei", *Proc. IEEE ICASSP*, pp. 2671-2674.

[33] Clermont, F. (1991), *Formant-Contour Models of Diphthongs: a Study in Acoustic Phonetics and Computer Modelling of Speech*, PhD Thesis, (Computer Sciences Laboratory, the Australian National University).

[34] Cole, R. A., Stern, R. M. and Lasry M. J. (1986), "Performing fine phonetic distinctions: Templates versus features", in Perkitt, J. S. and Klatt, D. M. (Eds.), *Variability and Invariance in Speech Processes* (Lawrence Erlbaum Assoc., Hillsdale, NJ), pp. 325-341.

[35] Cybenko, G. (1989), "Approximation by superpositions of a sigmoidal function", *Mathematics of Control, Signals, and Systems*, **2**(4), pp. 303-314.

[36] Dalsgaard, P. Andersen, O. and Barry, W. (1991), "Multi-lingual acoustic-phonetic features for a number of European languages", *Proc. Second European Conference on Speech Communication and Technology*, pp. 685-688.

[37] Dalsgaard, P. Andersen, O. Barry, W. and Jørgensen, R. (1992), "On the use of acoustic-phonetic features in interactive labelling of multi-lingual speech corpora", *Proc. IEEE ICASSP*, pp. I549–I552.

[38] Davis, K. H., Biddulph, R. and Balashek, S. (1952), "Automatic recognition of spoken digits", *J. Acoust. Soc. Am*, **24**, pp. 637-642.

[39] Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955), "Acoustic loci and transitional cues for consonants", *J. Acoust. Soc. Am.*, **27**, pp. 769-773.

[40] Denes, P. and Mathews, M. V. (1960), "Spoken digit recognition using time-frequency pattern-matching", *J. Acoust. Soc. Am*, **32**, pp. 1450-1455.

[41] Deng, L. and Erler, K. (1991), "Microstructural speech units and their HMM representation for discrete utterance speech recognition", *Proc. IEEE ICASSP*, pp. 193-196.

[42] Duda, R. O. and Hart, P. E. (1973), *Pattern Classification and Scene Analysis* (Wiley, New York).

[43] Dudley, H. and Balashek, S. (1958), "Automatic recognition of phonetic patterns in speech", *J. Acoust. Soc. Am*, **30**, pp. 721-732.

[44] Eimas, P. D., Siqueland, E. R., Jusczyk, P. and Vigorito, J. (1971), "Speech perception in infants", *Science*, **171**, pp. 303-306.

[45] Elenius, K. and Takács, G. (1990), "Acoustic–phonetic recognition of continuous speech by artificial neural networks", *Speech Transmission Lab. Quart. Progr. Status Report* (Royal Inst. Tech. Stockholm), **2-3**, pp. 1-44.

[46] Élman, J. L. (1988), "Finding structure in time", *CRL. Tech. UCSD, Tech. Report.*

[47] Erman, L. D. and Lesser, V. R. (1980), "The Hearsay-II speech understanding system: A tutorial", in Lea. W. A. (Ed.), *Trends in Speech Recognition* (Prentice-Hall, Inc., Englewood Cliffs, NJ), pp. 361-381.

[48] Fahlman, S. E. and Lebiere, C. (1990), "The cascade-correlation learning architecture", in Touretzky, D. (Ed.), *Advances in Neural Information Processing Systems 2* (Morgan Kaufmann, Los Altos, CA), pp. 524-532.

[49] Fant, C. G. M. (1957), "Analysis and synthesis of speech processes", in Malmberg, B. (Ed.), *Manual of Phonetics* (North-Holland Publishing Company, Amsterdam), pp. 173-277.

[50] Fant, C. G. M. (1960), *Acoustic Theory of Speech Production* (Mounton & Co. · 'S-Gravenhage, The Netherlands).

[51] Fant, C. G. M. (1966), "The nature of distinctive features", *Speech Transmission Lab. Quart. Progr. Status Report* (Royal Inst. Tech. Stockholm), **4**, pp. 1-15.

[52] Fant, C. G. M. (1969), "Distinctive features and phonetic dimensions", *Speech Transmission Lab. Quart. Progr. Status Report* (Royal Inst. Tech. Stockholm), **2-3**, pp. 1-18.

[53] Fant, C. G. M. (1973), *Speech Sounds and Features* (The MIT Press, Cambridge, MA).

[54] Fant, C. G. M. (1986), "Features: fiction and facts", in Perkell, J. and Klatt, D. H. (Eds.), *Invariance and Variability in Speech Processes* (Lawrence Erlbaum Assoc. Publishers, Hillsdale, NJ), pp. 481-492.

[55] Fant, C. G. M. (1990a), "Speech research in perspective", *Speech Communication*, **9**, pp. 171-176.

[56] Fant, C. G. M. (1990b), "The speech code.segmental and prosodic features", *Proc. ICSLP*, pp. 1389-1397.

[57] Flanagan, J. L. (1972), *Speech Analysis, Synthesis and Perception* (Springer-Verlag, Berlin).

[58] Foley, J. (1970), "Phonological distinctive features", *Folia Linguistica*, **4**, pp. 87-92.

[59] Forgie, J. W. and Forgie, C. D. (1959), "Results obtained from a vowel recognition computer program", *J. Acoust. Soc. Am*, **31**, pp. 1480-1489.

[60] Furui, S. and Sondhi, M. M. (Eds.)(1991) *Advances in Speech Signal Processing* (Marcel Dekker, Inc., New York).

[61] Glass, J. R. (1988), "Finding acoustic regularities in speech: Applications to phonetic recognition", *RLE Technical Report No. 536* (Research Laboratory of Electronics, MIT).

[62] Glass, J. and Zue, V. (1988), "Multi-level acoustic segmentation of continuous speech", *Proc. IEEE ICASSP*, pp. 429-432.

[63] Green, P. D. and Ainsworth, W. A. (1972), "Towards the automatic recognition of spoken basic English", *Machine Perception of Patterns and Pictures*, Inst. Phys., Conf. Ser. No. 13, pp. 161-168.

[64] Hajnal, A., Maass, W., Pudlak, P. Szegedy, M. amd Turan, G. (1987), "Threshold circuits of bounded depth", *Proc. of the 1987 IEEE Symposium on the Foundations of Computer Science*, pp. 99-110.

[65] Halle, M. (1964), "On the basis of phonology", In Fodor, J. A. and Katz, J. J. (Eds.), *The Structure of Language: Readings in the Philosophy of Language* (Prentice-Hall, Englewood Cliffs, NJ), pp. 324-333.

[66] Hanson, G. (1967), "Dimensions in speech sound perception: An experimental study of vowel perception", *Ericsson Tech. Report*, **23**, pp. 3-175.

[67] Hopfield, J. J. (1982), "Neural networks and physical systems with emergent collective computational abilities", *Proc. Natl. Acad. Sci. USA*, **79**, pp. 2554-2558.

[68] Howard, I. S. and Huckvale, M. A. (1989), "Two-level recognition of isolated word using neural nets", *First International Conference on Artificial Neural Networks*, pp. 90-94.

[69] Huang, X. D. and Jack, M. A. (1989), "Semi-continuous hidden Markov models for speech recognition", *Computer, Speech and Language*, **3**(3), pp. 239-252.

[70] Huang, X. D., Ariki, Y., Jack, M. A. (1990), *Hidden Markov Models for Speech Recognition* (Edinburgh University Press, Edinburgh).

[71] Huckvale, M. (1990), "Exploiting speech knowledge in neural nets for speech recognition", *Speech Communication*, **9**, pp. 1-13.

[72] Huckvale, M. and Howard, I. (1990), "Phonetic feature analysis for a monosyllabic word recognition task", *Proc. Institute of Acoustic Conference on Speech and Hearing.*

[73] Hush, D. R. and Horne, B. G. (1993), "Progress in supervised neural networks", *IEEE Signal Processing Magazine*, January, 1993, pp. 8-39.

[74] Hyman, L. M. (1975), *Phonology: Theory and Analysis* (Holt, Rinehart and Winston, USA).

[75] Jakobson, R. (1938), "Observations sur le classement phonologique des consonnes", *Proc. 3rd Int. Congr. Phon. Sci.*, pp. 34-41.

[76] Jakobson, R., Fant, C. G. M. and Halle, M. (1952), "Preliminaries to speech analysis: the distinctive features and their correlates", *Technical report*, **13**, (Acoustic Laboratory, MIT).

[77] Jakobson, R. and Halle, M. (1956), *Fundamentals of Language* (Mounton & Co. · 'S-Gravenhage, The Netherlands).

[78] Jakobson, R., Fant, C. G. M. and Halle, M. (1961), *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates* (MIT Press, Cambridge, MA).

[79] Jakobson, R. and Waugh, L. (1979), *The Sound Shape of Language* (Harvester Press, Great Britain).

[80] Jelinek, F. (1976), "Continuous speech recognition by statistical methods", *Proc. IEEE*, **64**(4), pp. 532-556.

[81] Jones, D. (1918), *An Outline of English Phonetics* (W. Heffer & Sons Ltd., Cambridge, MA).

[82] Jones, D. (1956), *An Outline of English Phonetics* (W. Heffer & Sons Ltd., Cambridge, MA).

[83] Jordan (1986), "Serial order: A parallel distributed processing approach", *UCSD, Tech. Report 8604.*

[84] Juang, B. H. and Rabiner, L. R. (1985), "Mixture autoregressive hidden Markov models for speech signals", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **33**, pp. 1404-1413.

[85] Kewley-Port, D. (1979), "Spectral continuity of burst and formant transitions as cues to place of articulation in stop consonants", Speech communication papers presented at the *97th meeting of the Acoustical Society of America*, pp. 175-178.

[86] Kewley-Port, D. (1982), "Measurement of formant transitions in naturally produced stop consonant-vowel syllables", *J. Acoust. Soc. Am.*, **72**, pp. 379-389.

[87] Kewley-Port, D. (1983), "Time-varying features as correlates of place articulation in stop consonants", *J. Acoust. Soc. Am.*, **73**, pp. 322-335.

[88] Klatt, D. H. (1968), "Structure of confusions in short term memory between English consonants", *J. Acoust. Soc. Am.*, **44**, pp. 401-407.

[89] Klatt, D. H. (1979), "Speech perception: A model of acoustic-phonetic analysis and lexical access", *J. Phonetics*, **7**, pp. 279-312.

[90] Kohonen, T. (1988), "The 'neural' phonetic typewriter", *IEEE Computer*, March, pp. 11-22.

[91] Kohonen, T. (1989), *Self Organization and Associative Memory* (Third edition), (Springer-Verlag, Berlin).

[92] Kohonen, T., Mäkisara, K. and Saramäki, T. (1984), "Phonotopic Maps — insightful representation of phonological features for speech recognition", *Proc. 7th ICPR*, pp. 182-185.

[93] Kohonen, T., Barna, G. and Chrisley, R. (1988), "Statistical pattern recognition with neural networks: Benchmarking studies", *IEEE Proc. ICNN*, **1**, pp. 61-68.

[94] Ladefoged, P. (1971), *Preliminaries to Linguistic Phonetics* (University of Chicago Press, Chicago, IL).

[95] Ladefoged, P. (1975a), *A Course in Phonetics* (First edition), (Harcourt Brace Jovanovich Publishers, New York).

[96] Ladefoged, P. (1975b), *Three Areas of Experimental Phonetics* (Fourth edition), (Oxford University Press, London).

[97] Ladefoged, P. (1982), *A Course in Phonetics* (Second edition), (Harcourt Brace Jovanovich Publishers, New York).

[98] Lahiri, A. (1980), "Coronal stops in Malayalam", *Brown University Working Papers*, **4**, pp. 81-95.

[99] Lahiri, A. and Blumstein, S. E. (1981), "A reconsideration of acoustic invariance in stop consonants: Evidence from cross-language studies", *J. Acoust. Soc. Am. Suppl.*, 1 **70**, S39.

[100] Lahiri, A., Gewirth, L. and Blumstein, S. E. (1984), "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-languages study", *J. Acoust. Soc. Am.*, **76**, pp. 391-404.

[101] Le Cun, Y. (1985), "Une procédure dàpprentissage pour réseaux à seuil asymétrique", *Proceedings Cognitiva'85*, pp. 599-604.

[102] Le Cun, Y., Denker, J. S. and Solla, S. A. (1990), "Optimal brain damage.", In Touretzky, D. (Ed.) *Advances in Neural Information Processing Systems 2* (Morgan Kaufmann, Los Altos, CA), pp. 598-605.

[103] Lea, W. A. (1973), "An approach to syntactic recognition without phonemics", *IEEE Trans. on Audio and Electroacoustics*, **AU-21**, pp. 249-258.

[104] Lea, W. A. (Ed.) (1980), *Trends in Speech Recognition* (Prentice-Hall, Inc., Englewood Cliffs, NJ).

[105] Lee, F. K., Hon, W. H., Hwang, M. Y., Mahajan, S. and Reddy, R. (1989), "The Sphinx speech recognition system", *Proc. IEEE ICASSP*, pp. 445-448.

[106] Lee, F.K. (1990), "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, April, 1990.

[107] Lehiste, I. (1967), *Readings in Acoustic Phonetics* (MIT Press, Cambridge, MA).

[108] Lesser, V. R., Fennell, R. D., Erman, L. D. and Reddy, D. R. (1975), "Organization of the Hearsay-II speech understanding system", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **23**, pp. 11-23.

[109] Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983), "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", *Bell Syst. Tech. J.* , **62**, pp. 1035-1074.

[110] Levinson, S. E. and Roe, D. B. (1990), "A perspective on speech recognition", *IEEE Communication Magazine*, January, 1990.

[111] Liberman, A. M., Delattre, P. C., Cooper, F. S. and Gerstman, L. J. (1954), "The role of consonant–vowel transitions in the perception of the stop and nasal consonants", *Psychol. Mono.*, **68**, pp. 1-13.

[112] Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. (1957), "The discrimination of speech sounds within and across phoneme boundaries", *J. Exptl. Psychol.*, **54**, pp. 358-368.

[113] Liberman, A. S. (1974), "The order of rules in phonology and the reality of distinctive features", *Linguistics*, **126**, pp. 45-62.

[114] Lieberman, P. and Blumstein S. E. (1988), *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge University Press, Cambridge, MA).

[115] Lindau, M. and Ladefoged, P. (1986), "Variability of feature specifications", in Perkell, J. and Klatt, D. H. (Eds.), *Invariance and Variability in Speech Processes* (Lawrence Erlbaum Assoc. Publishers, Hillsdale, NJ), pp. 464-478.

[116] Lippmann, R. P. (1987), "An introduction to computing with neural nets", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **4**(2), pp. 4-22.

[117] Lippmann, R. P. (1989), "Review of neural networks for speech recognition", *Neural Computation*, **1**(1), pp. 1-38.

[118] Lowerre, B. T. (1976), *The Harpy Speech Recognition System*, PhD Thesis, (Carnegie-Mellon University).

[119] Lowerre, B. T. and Reddy, D. R. (1980), "The Harpy speech understanding system", in Lea, W. A. (Ed.), *Trends in Speech Recognition* (Prentice-Hall, Inc., Englewood Cliffs, NJ), pp. 340-360.

[120] Makhoul, J. and Schwartz, R. (1986), "Ignorance modeling", in Perkell, J. S. and Klatt, D. M. (Eds.), *Variability and Invariance in Speech Processes* (Lawrence Erlbaum Assoc., Hillsdale, NJ), pp. 344-345.

[121] Makhoul, J., El-Jaroudi, A. and Schwartz, R. (1989), "Formation of disconnected decision regions with a single hidden layer", *Proc. of International Joint Conference on Neural Networks*, **1**, pp. 455-460.

[122] Mariani J. (1989), "Recent advances in speech processing", *Proc. IEEE ICASSP*, pp. 429-440.

[123] Markov A. A. (1913), "An example of statistical investigation in the text of Eugen Onyegin illustrating coupling of tests in chains", *Proc. Acad. Sci. St Petersburgh VI ser.*, **7**, pp. 153-162.

[124] McCarthy, J. J. (1988), "Feature geometry and dependency: A review", *Phonetica*, **45**, pp. 84-108.

[125] McDermott, E. Iwamida, H. Katagiri, S. and Tohkura, Y. (1990), "Shift-tolerant LVQ and Hybrid LVQ-HMM for phoneme recognition", in Waibel, A. and Lee, K-F. (Eds.), *Readings in Speech Recognition* (Morgan and Kaufmann Publishers, Inc., San Mateo, CA).

[126] Mercier, G. Nouhen, A. Quinton, P. and Siroux, J. (1980), "The Keal speech understanding system", in Simon, J. C. (Ed.), *Spoken Language Generation and Understanding* (D. Reidel Pub. Co. Holland), pp. 525-543.

[127] Mercier, G. D., Bigorgne, L. Miclet, L., Guennec, L. and Querre, M. (1989), "Recognition of speaker-dependent continuous speech with Keal", *IEE Proceedings-I*, **136**(2), pp. 145-154.

[128] Miller, G. and Nicely, P. E. (1955), "An analysis of perceptual confusions among English consonants", *J. Acoust. Soc. Am.*, **27**, pp. 338-352.

[129] Minsky, M. and Papert, S. (1969), *Perceptrons: An Introduction to Computational Geometry* (MIT Press, Cambridge, MA).

[130] Mitchell, A. G. (1946), *The Pronunciation of English in Australia* (Halstead Press, Sydney).

[131] Mitchell, A. G. (1962), *Spoken English* (St Martin's Press, New York).

[132] Mitchell, A. G. and Delbridge A. (1965), *The Pronunciation of English in Australia* (Halstead Press, Sydney).

[133] Morgan, N. and Bourlard, H. (1990), "Continuous speech recognition using multilayer perceptrons with hidden Markov models", *Proc. IEEE ICASSP*, pp. 413-416.

[134] Mozer, M. C. (1988), "A focused back-propagation algorithm for temporal pattern recognition", *Tech. Report CRG-TR-88-3* (University of Toronto), 1988.

[135] Myers, C. S. and Rabiner, L. R. (1981), "Connected digit recognition using a level building DTW algorithm", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **29**(3), pp. 351-363.

[136] Ney, H. (1984), "The use of a one-stage Dynamic Programming algorithm for connected word recognition", *IEEE Trans. on Acoustics, Speech, and Signal Processing* **32**(2), pp. 263-271.

[137] Nieman, H. (1982), "The Erlagen system for the recognition and understanding of continuous German speech", *Informatik Fachberichte*, **57**, pp. 330-348.

[138] Nilsson, N. J.(1965), *Learning Machines: Foundations of Trainable Pattern Classifying Systems* (McGraw-Hill, New York).

[139] Oasa, H. (1980), *A Quantitative Study of Regional Variations in Australian English*, Masters Thesis (Dept. of Linguistics, the Australian National University).

[140] Oshika, B. T., Zue, V. W., Weeks, R. V., Neu, H. and Aurbach, J. (1974), "The role of phonological rules in speech understanding research", *IEEE Symp. Speech Recognition*, pp. 204-207.

[141] Oshika, B. T. and Zue, V. W. (1975), "The role of phonological rules in speech understanding research", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **23**, pp. 104-112.

[142] Parker, D. B. (1985), "Learning logic", *Tech. Report, TR-47*, (cerems, MIT).

[143] Paul, D. B, Lippmann, R. P., Chen, Y. and Weinstein, C. (1986), "Robust HMM-based techniques for recognition of speech produced under stress and in noise", *Proceedings of Speech Tech.*

[144] Peeling, S. M. and Moore, R. K., (1988), "Isolated digit recognition experiments using the multilayer perceptron", *Speech Communication*, **7**(4), pp. 403-410.

[145] Perkell, J. S. and Klatt, D. M. (Eds.) (1986), *Variability and Invariance in Speech Processes* (Lawrence Erlbaum Assoc., Hillsdale, NJ).

[146] Pols, L. C. W. (1986), "Variation and interaction in speech", in Perkell, J. and Klatt, D. H. (Eds.), *Invariance and Variability in Speech Processes* (Lawrence Erlbaum Assoc. Publishers, NJ), pp. 140-154.

[147] Poritz, A. B. and Richter, A. G. (1986), "On hidden Markov models in isolated word recognition", *Proc. IEEE ICASSP*, pp. 705-708.

[148] Potter, R. K., Kopp, G. A. and Green, H. C. (1947), *Visible Speech* (D. Van Nostrand Company, Inc. New York).

[149] Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G. (1979), "Speaker-independent recognition of isolated words using clustering techniques", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **27**(4), pp. 336-349.

[150] Rabiner, L. R. and Levinson, S. E. (1981), "Isolated and connected word recognition – theory and selected applications", *IEEE Trans. on Communications*, **29**(5), pp. 621-659.

[151] Rabiner, L. R. Juang, B. H., Levinson, S. E., and Sondhi, M. M. (1985), "Recognition of isolated digits using hidden Markov models with continuous mixture densities", *AT&T Technical Journal*, **64**, pp. 1211-1234.

[152] Rabiner, L. R. and Juang, B. H. (1986), "An introduction to hidden Markov models", *IEEE ASSP Magazine*, pp. 4-16.

[153] Rabiner, L. R. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, pp. 257-286.

[154] Rabiner, L. R., Wilpon, J. G., and Soong, F. K. (1989), "High performance connected digit recognition using hidden Markov models", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **37**(8), pp. 1214-1225.

[155] Ran, S. and Millar, J. B. (1990), "Exploring the phonetic structure of speech signal using multi-layer perceptrons", *Proc. of the Third Australian International Conference on Speech Science and Technology*, pp. 22-27.

[156] Ran, S. and Millar, J. B. (1991a), "Phoneme discrimination using hierarchically organised connectionist networks", *Proc. Second Australian Conference on Neural Networks*, pp. 279-282.

[157] Ran, S. and Millar, J. B. (1991b), "Phoneme classification using neural networks based on acoustic-phonetic structure", *Proc. Second European Conference on Speech Communication and Technology*, pp. 125-132.

[158] Ran, S. and Millar, J. B. (1991c), "Acoustic-phonetic structure analysis for phoneme identification using neural networks", presented at *Eighth Australian Language and Speech Conference.*

[159] Ran, S. and Millar, J. B. (1992), "Phonetic feature extraction using artificial neural networks", *Proc. Fourth Australian International Conference on Speech Science and Technology*, pp. 22-27.

[160] Ran, S. and Millar, J. B. (1993a), "Two schemes of phonetic feature extraction using artificial neural networks", *Proc. Third European Conference on Speech Communication and Technology*, pp. 1607-1610.

[161] Ran, S. and Millar, J. B. (1993b), "Phoneme recognition based on phonetic features usng artificial neural networks", *Proc. International Conference on Signal Processing-93*, pp. 663-666.

[162] Reddy, D. R., Erman, L. D., and Neely, R. B. (1973), "A model and a system for machine recognition of speech", *IEEE Trans.*, **AU-21**, pp. 229-238.

[163] Reenen, P. V. (1982), *Phonetic Feature Definitions: Their Integration Into Phonology and Their Relation to Speech: A Case Study of the Feature NASAL* (Foris Publications, The Netherlands).

[164] Richard, M. D. and Lippmann, R. P. (1991), "Neural network classifiers estimate Bayesian a posteriori probabilities", *Neural Computation*, **3**(4), pp. 461-483.

[165] Robinson, A. J. (1989), *Dynamic Error Propagation Networks*, PhD Thesis, (Cambridge university).

[166] Rosenblatt, R. (1958), "The Perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, **65**, pp. 386-408.

[167] Rosenblatt, R. (1959), *Principles of Neurodynamics* (Spartan Books, New York).

[168] Rumelhart D. E., Hinton G. E. and McClelland J. L. (1986a), "A general framework for parallel distributed processing", in Feldman, J. A., Hayes, P. J. and Rumelhart, D. E (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (MIT Press, Cambridge, MA), pp. 45-76.

[169] Rumelhart D. E., Hinton G. E. and Williams, R. J. (1986b), "Learning internal representations by error propagation", in Feldman, J. A., Hayes, P. J. and Rumelhart, D. E (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (MIT Press, Cambridge, MA), pp. 318-362

[170] Sagey, E. (1990), *The Representation of Features in Non-Linear Phonology: the Articulator Node Hierarchy* (Garland Publishing Inc., New York).

[171] Sakoe, H. and Chiba, S. (1978), "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **26**, pp. 43-49.

[172] Sakoe, H. (1979), "Two level DP-matching — A dynamic programming-based pattern matching algorithm for connected word recognition", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **27**(6), pp. 588-595.

[173] Sakoe, H., Isotani, R. Yoshida, K. Iso, K. and Watanabe, T. (1989), "Speaker independent word recognition using dynamic programming neural networks", *Proc. IEEE ICASSP*, pp. 29-32.

[174] Shepard, R. N. (1972), "Psychological representation of speech sounds", Edward, E., David, Jr. and Denes, P. B. (Eds.), *Human Communication: A unified View* (McGraw-Hill, New York), pp. 67-113.

[175] Shoup, J. E. (1980), "Phonological aspects of speech recognition", in Lea, W. A. (Ed.), *Trends in Speech Recognition* (Prentice-Hall, Inc, Englewood Cliffs, NJ).

[176] Singh, S. and Black, J. W. (1966), "Study of twenty-six intervocalic consonants as spoken and recognised by four language groups", *J. Acoust. Soc. Am.*, **39**, pp. 372-387.

[177] Singh, S. and Woods, D. R. (1971), "Perceptual structure of 12 American English vowels", *J. Acoust. Soc. Am.*, **49**, pp. 1861-1866.

[178] Singh, S. Woods, D. R. and Becker, G. M. (1972), "Perceptual structure of 22 prevocalic English consonants", *J. Acoust. Soc. Am.*, **52**, pp. 1698-1713.

[179] Singh, S. (1976), *Distinctive Features: Theory and Validation* (University Park Press, Baltimore, MD).

[180] Singh, S. and Singh, K. S. (1976), *Phonetics: Principles and Practices* (University Park Press, Baltimore, MD).

[181] Stern, P. E., Eskenazi, M. and Memmi, D. (1986), "An expert system for speech spectrogram reading", *Proc. IEEE ICASSP*, pp. 1193-1196.

[182] Stevens, K. N. and House, A. S. (1956), "Studies of formant transitions using a vocal tract analog", *J. Acoust. Soc. Am.*, **28**, pp. 578-585.

[183] Stevens, K. N. House, A. S. and Paul, A. P. (1966), "Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation", *J. Acoust. Soc. Am.*, **40**, pp. 123-132.

[184] Stevens, K. N. Liberman, A. M., Studdert-Kennedy, M. and Öhman, S. (1969), "Crosslanguage study of vowel perception", *Language and Speech*, **12**, pp. 1-23.

[185] Stevens, K. N. (1972), "The quantal nature of speech: Evidence from articulatory-acoustic data", in David, E. E. and Denes, P. B. (Eds.), *Human Communications: A Unified View* (McGraw-Hill, New York), pp. 67-113.

[186] Stevens, K. N. and Blumstein, S. E. (1978), "Invariant cues for place of articulation in stop consonants", *J. Acoust. Soc. Am.*, **64**, pp. 1358-1368.

[187] Stevens, K. N. (1980), "Acoustic correlates of some phonetic categories", *J. Acoust. Soc. Am.*, **68**(3), pp. 836-842.

[188] Stevens, K. N. (1986), "Models of phonetic recognition II: An approach to feature-based recognition", *Proc. of Montreal Symposium on Speech Recognition.*

[189] Stevens, K. (1992), "Lexical access from features", *Speech Communication Group Working Papers*, **8**, (Research Laboratory of Electronics, MIT).

[190] Studdert-Kennedy, M. and Schankweiler, D. (1970), "Hemispheric specialization for speech perception", *J. Acoust. Soc. Am.*, **48**, pp. 579-594.

[191] Suomi, K. (1985), "The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables", *Journal of Phonetics*, **13**, pp. 267-285.

[192] Suomi, K. (1987), "On spectral coarticulation in stop-vowel-stop syllables: implications for automatic speech recognition", *Journal of Phonetics*, **15**, pp. 85-100.

[193] Sussman, H. M., McCaffrey, H. A. and Matthews, S. A. (1991), "An investigation of locus equations as a source of relational invariance for stop place categorization", *J. Acoust. Soc. Am.*, **90**(3), pp. 1309-1325.

[194] Tappert, C. C. (1974), "Experiments with a tree search method for converting noisy phonetic representation into standard orthography", *IEEE Symp. Speech Recognition*, pp. 261-266.

[195] Terbeek, D. and Harshman, R. (1971), "Crosslanguage differences in the perception of natural vowel sounds", *UCLA Working Papers in Phonetics.*

[196] Teuber, H-L. (1976), "The brain and human behavior", *Proc. 21st International Congress of Psychology*, pp. 119-163.

[197] Trubetzkoy, N. S. (1939), "Grundzüge der Phonologie", *Travaux du Cercle linguistique de Prague 7*, (English translation by Baltaxe, C.: *Principles of Phonology*, University of California Press, Berkeley, CA, 1969).

[198] Velichko, V. M. and Zagoruyko, N. G. (1970), "Automatic recognition of 200 words", *Int. J. Man-Machine Studies*, **2**, pp. 223-234.

[199] Vennemann, T. (1972), "On the theory of syllabic phonology", *Linguistische Berichte*, **18**, pp. 1-18.

[200] Vintsyuk, T. K. (1971), "Element-wise recognition of continuous speech composed of words from a specified dictionary", *Kibernetika*, **7**, pp. 133-143.

[201] Viterbi, A. J. (1967), "Error bounds for convolutional codes and asymptotically optimum decoding algorithm", *IEEE Trans. on Information Theory*, **13**(2), pp. 260-269.

[202] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1987), "Phoneme recognition using time-delay neural networks", *ATR Research Report TR-I-0006.*

[203] Waibel, A., Sawai, H. and Shikano, K. (1989), "Consonant recognition by modular construction of large phonemic time-delay neural networks", *Proc. IEEE ICASSP*, pp. 112-115.

[204] Waibel, A. and Lee, K–F. (Eds.) (1990), *Readings in Speech Recognition*, (Morgan Kaufmann Publishers, Inc., San Mateo, CA).

[205] Wang, W. S.–Y. (1967), "Phonological features of tone", *International Journal of American Linguistics*, **33**, pp. 93-105.

[206] Wang, M. and Bilger, R. C. (1973), "Consonant confusions in noise: A study of perceptual features", *J. Acoust. Soc. Am.*, **54**, pp. 1248-1266.

[207] Watrous, R. L., Shastri, L. and Waibel, A. H. (1987), "Learned phonetic discrimination using connectionist networks", *Proc. European Conference on Speech Technology*, pp. 377-380.

[208] Werbos, P. (1974), *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, PhD Thesis (Harvard University).

[209] White, H. (1989), "Learning in artificial neural networks: A statistical perspective", *Neural Computation*, **1**(4), pp. 425-464.

[210] Wickelgren, W. A. (1966), "Distinctive features and errors in short-term memory for English consonants", *J. Acoust. Soc. Am.*, **39**, pp. 388-398.

[211] Williams R. J. (1986), "The logic of activation functions", in Feldman, J. A., Hayes, P. J. and Rumelhart, D. E (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (The MIT Press, Cambridge, MA), pp. 423-443.

[212] Wiren, J. and Stubbs, H. L. (1956), "Electronic Binary selection system for phoneme classification", *J. Acoust. Soc. Am.*, **28**, pp. 1082-1091.

[213] Woods, W. A. (1974), "Motivation and overview of SPEECHLIS: an experimental prototype for speech understanding research", *IEEE Trans. on Acoustics, Speech and Signal Processing*, **23**, pp. 2-10.

[214] Zue, V. W. and Schwartz, R. M. (1980), "Acoustic processing and phonetic analysis", in Lea, W. A. (Ed.), *Trends in Speech Recognition* (Prentice-Hall, Inc, Englewood Cliffs, NJ).

[215] Zue, V. W. (1985), "The use of speech knowledge in automatic speech recognition", *Proc. of IEEE*, **73**(11), pp. 1602-1615.

[216] Zue, V. W. and Lamel, L. F. (1986), "An expert spectrogram reader: A knowledge-based approach for speech recognition", *Proc. IEEE ICASSP*, pp. 1197-1200.

[217] Zue, V. W. (1988), "Recent speech recognition results at MIT", *DARPA Review Meeting.*

[218] Zue, V. Glass, J. Phillips, M. and Seneff, S. (1989), "Acoustic segmentation and phonetic classification in the Summit system", *Proc. IEEE ICASSP*, pp. 389-392.